Active Action Recognition in Humans and Robots

Carmelo Calafiore

A Thesis Submitted for the Degree of Doctor of Philosophy

Department of Psychology

and

School of Computer Science and Electronic Engineering

University of Essex

February 2024

Acknowledgements

I am very proud of myself for submitting this PhD thesis. However, it was a very long and difficult journey for me, I faced a lot of difficult challenges, and I made a lot of mistakes on the way to this submission. I could not have completed this thesis without the support and encouragement I have received throughout my academic journey.

I thank my ex-schoolteacher, Mr. Calogero Castellana, for making me believe in myself when I was too young and naïve. I used to be a very bad student in school, and I did not have any interest in studying. I still remember I used to say to people and myself "I cannot be a psychologist" because "I don't know how to study", "I don't like to study", "I don't want to learn how to study" and "it's too late for me". One day, Mr. Castellana asked all of us what job we wanted to do in life. I said I wanted to be a psychologist, but I could not do it because I was not a good student. I do not really remember what he said to me that day, but I remember he talked to me like I could make it and it was not too late for me. Something triggered me that day and made me decide to give it a go.

I thank my PhD supervisors, Dr Tom Foulsham and Dr Dimitri Ognibene, for choosing me as their PhD student among many other valuable candidates, and for being very patient in helping me for such a long time. My supervisors guided and corrected me with their irreplaceable expertise until the very end. Sometimes, they put some pressure on me, but that was for goodwill to make me more productive.

I thank my examiners, Dr Paul Hibbard and Dr Dietmar Heinke, for their priceless technical suggestions on how to improve my thesis. They had to give up some of their valuable time to read my thesis, assess it, travel and examinate me for three straight hours for my viva.

I thank the University of Essex for funding my research for the first three years and providing unlimited access to the High Performance Computing Facility (Ceres) cluster for my deep learning experiments. The Ceres cluster is a computational cluster comprising 2192 processors (including multi-threading), 43.5Tb total RAM, 24 GPUs, and 660Tb of dedicated storage. Special thanks go to the cluster maintainer, Mr. Stuart Newman, for his assistance in fixing my issues with the cluster.

I thank my parents, Mr. Angelo Calafiore and Mrs. Assunta Alfonsa D’Amore. Even if they are not highly educated, they supported and funded my academic journey from the beginning to the end. I hope that I have made them proud.

Last but not least, I thank my girlfriend, Miss Xiangshu Huang, for understanding the times when I was unable to spend time with her, for supporting me financially for half a year during the difficult times, and for trusting me until the end. I am endlessly appreciative of her.

Abstract

Active action recognition is the selection of the best viewpoints for more accurate and faster action recognition. The studies in this thesis aimed to examine whether and how humans and robots can process efficient active action recognition.

Participants could either be humans or robots. The robots were recurrent convolutional neural networks trained to classify actions using supervised learning and select the next best view by deep Q-learning. Each participant classified human actions from different viewpoints in either active or passive conditions. The participants in the active condition could select the viewpoint movements, whereas those in the passive conditions had no control over their viewpoints. The passive conditions could either be no view movement (NM) or random view movement (RM). In the NM condition, the view did not change within trials. In the RM condition, the viewpoint moved randomly.

The studies showed that humans were slightly more accurate and faster in recognizing actions in the active condition than in the RM condition. However, some studies did not replicate this advantage in humans. Nevertheless, the robots were more accurate in the active condition than in the passive conditions.

The efficient viewpoints for action recognition of humans and robots were the ones from which their action recognition was more accurate or faster in their NM condition. The efficient views tended to be the top, the front, and the side views with respect to the actors.

Humans in the active condition selected the efficient viewpoints more often than the others. However, robots in the active condition did not choose the efficient viewpoints more frequently than the others. Instead, the robots moved their viewpoints far from their starting positions, suggesting they learned to get more accurate action classifications by

observing actions from many viewpoints at different timepoints rather than from a few efficient viewpoints.

Table of Contents

1 Introduction to Active Action Recognition

Accurate and fast action recognition is essential for human and robotic observers to understand the world, appropriately react to it and ultimately fulfil their needs and goals. Indeed, in human observers, action recognition can ease object (actor) recognition (Blasing & Sauzet, 2018; Mitchell & Curry, 2016), as well as object recognition can facilitate action recognition (Ferstl et al., 2017; Knoblich & Sebanz, 2006; Schutz-Bosbach et al., 2006; Sebanz et al., 2006). Furthermore, recognising actions facilitate human observers to detect the intentions, feelings (Foulsham & Lock, 2015) and thoughts of the actors. In the last decade, robotic observers (which, sometimes, in this thesis, I may call robotic vision models, computer vision models, computer models, robotic participants or robots) have had a large success by deep neural networks (NNs) in action recognition (Aghaei et al., 2021; Al-Faris, Chiverton, Ndzi, et al., 2020; Al-Faris, Chiverton, Yang, et al., 2020; Dai et al., 2020; Donahue et al., 2015; Li et al., 2018; Majd & Safabakhsh, 2020; Ng et al., 2015). However, there is an overall lack of research regarding how human and robotic observers may be capable of efficient active action recognition. Efficient active action recognition is the skill of some observers to efficiently select the viewpoints for more accurate or faster action recognition. Therefore, the main aim of this thesis is to discover whether and how human and robotic observers are capable of efficient active action recognition.

Most studies about action recognition in humans and robotic vision models are actually about the passive action recognition because the viewpoints of most experiments studying action recognition are fixed and predefined and the observers cannot move their own view positions away from the inefficient viewpoints to the efficient ones. The efficient viewpoints for some observers are the ones from where action recognition is more accurate or faster than the inefficient view positions. This implies that efficient views can only be without-obstacle views, while inefficient views can either be with-obstacles or without-obstacles views. The with-obstacle views are the ones from which there are some non-transparent obstacles, such as a wall, occluding the sight of the action and its participants, whereas the without-obstacle views are the ones from where there is nothing, except for transparent air and windows, covering the vision of the action. Therefore, most studies do not take in consideration that, in the real world, actions can be either be clearly visible from the efficient views or occluded or ambiguous from the inefficient view positions. For instance, a police officer (observer) can clearly see a thief (actor) stealing items on sale from a shop shelf with their right hand from some views (efficient views) because there is nothing between these viewpoints (without-obstacle views) and the stealing hand, while the police officer cannot clearly see this hand from some other positions (inefficient viewpoints) because a wall (non-transparent obstacle) is between the hand and these viewpoints (with-obstacle views). From the inefficient with-obstacle views, a passive (non-moving) observer may just guess whether the potential thief is stealing, while an active observer, like the police officer, can move their view away from the inefficient with-obstacle views to some efficient without-obstacle positions and correctly detect the crime. Therefore, active action recognition is essential for intelligent observers to cope with action occlusion and action ambiguity and improve the accuracy and the speed of their action recognition.

Active action recognition (or active action classification) is a complex skill that requires at least four visual skills. Let me clarify here that the words recognition and classification are synonyms in the whole thesis. These are object classification, action classification, visual perspective taking and on-line egocentric localization. Object classification (or, at least, object perception) may generally support action classification because actions can only be seen on objects. In fact, actions are changes of states made by some effectors to some recipients by some instruments. Effectors, recipients and instruments are the actors or the agents of the actions. Since effectors, recipients and instruments of an action are either animate or inanimate objects, then classifying the objects that participate in the actions can ease action classification (Ferstl et al., 2017; Knoblich & Sebanz, 2006; Schutz-Bosbach et al., 2006; Sebanz et al., 2006). Obviously, action classification is vital for active action classification. By definition, active action classification is action classification with efficient viewpoint navigation to quickly reach the most efficient viewpoints where to clearly look at the action from.

The last two visual skill of active action classification are crucial for efficient viewpoint navigation. Efficient view navigation involves visual perspective taking of the actors to localize efficient viewpoints with respect to the actors. It also requires visual perspective taking of the viewpoints to predict how the action can be seen from them, discriminate the with-obstacle and without-obstacle viewpoints and exclude the with-obstacle views. Finally, real-time egocentric localization of the selected viewpoint is heavily important to guide fast viewpoint navigation to the target views in real time. The egocentric localization must update the egocentric coordinates of the target views at high refresh rate such that they are always accurate at any moment as the observer quickly navigates to the target view.

1.1 Object Classification

Object classification or object recognition is the ability of intelligent observers of matching the perceived objects with their memorized classes of objects. The classes of objects are finite representations of the observers that describe and group the infinite possible objects of the world with similar features. The features of the objects include shapes, colours, sizes, weights and even potential class of actions. The classes of objects and their defining proprieties are stored in the long-term memory of the observers. Objects are all physical entities of the world. Objects can either be animated such as people, dogs and birds or inanimate like tables, houses and apples. Every object can further be in different possible states like location, orientation, light condition and executing action class.

There are infinite possible objects which can be at infinite possible states. The objects are infinite because they can have infinite possible combinations of different shapes, colours, sizes, weights and potential actions. The shapes, colours, sizes, weights and potential actions are infinite because they can be defined by continuous variables in the real world such as lengths and orientations of their edges (shape), reflected light wave frequencies and intensities from their surfaces (colours), body volume (size), body weight and locations. By definition, continuous variables (for instance, length) contain infinite values (1.0001 mm, 1.00001 mm, 1.000001 mm, …) within any tiny range of values (1.0 mm < x < 1.001 mm) regardless of how tiny that range is (.001 mm). Thus, there are infinite possible objects because they can be defined by infinite values. Furthermore, what makes it even more challenging for the observers is that every object can further be in infinite possible states. The states of every object are infinite because its location, orientation, light condition and executing action are also defined by continuous variables. However, intelligent observers classify infinite possible objects at infinite possible states into finite possible object classes.

1.1.1 Bayesian Inferences in Object Classification

Numerous studies (Battaglia et al., 2003; Friston, 2003; Friston et al., 2006; Kording & Wolpert, 2004; Meijer et al., 2019) have provided evidence which suggest that the human brain are complicated functions that makes Bayesian estimations of the hidden proprieties of the world given some sensory information. Thus, by assuming that human brain and NNs are Bayesian estimators of the properties of the world given some observations, let me describe how human and robotic observers would classify objects given some observations by Bayesian inferences. The Bayesian observers classify objects, by computing the posterior probability of the object classes given some observations of the objects. Firstly, let me define the observation s_t as the t^th image of the video v with T images showing an instance of the i^th object class o_i doing an instance of the j^th action class a_j.

$$s_t \in v \tag{1.1}$$

where t is an integer in the range 0 ≤ t < T and

$$v = \{s_0, s_1, \dots, s_{T-1}\} \tag{1.2}$$

For O possible object classes where O is a positive integer, the posterior probability $P(o_i|s_t)$ of the i^th object class o_i given the t^th image s_t of the video v is defined in Equation 1.3.

$$P(o_i|s_t) = \frac{P(s_t|o_i) \times P(o_i)_t}{P(s_t)} \tag{1.3}$$

i is any integer in the range 0 ≤ i < O. $P(i_t|o_i)$ is the conditional probability of the image s_t given the object class o_i. $P(o_i)_t$ is the prior probability of the object class o_i at the timepoint t regardless of the observed image s_t. Equation 1.4 shows that $P(o_i)_t$ is equal to the prior probability $P(o_i)$ for the first timepoint with t=0. For t>0, the prior $P(o_i)_t$ at timepoint t is updated to the previous posterior probability $P(o_i|s_{t-1})$ of the object class o_i given the previous observation s_{t-1}.

$$P(o_i)_t = \begin{cases} P(o_i), & if\ t = 0 \\ P(o_i|s_{t-1}), & else\ if\ t > 0 \end{cases} \qquad 1.4$$

The observer evidence $P(s_t)$ of the image s_t in Equation 1.3 is defined as:

$$P(s_t) = \sum_{u=0}^{O-1} P(s_t|o_u) \times P(o_u)_t \qquad 1.5$$

where $P(o_u)_t$ can be defined by replacing all i with u in Equation 1.4.

1.2 Action Classification

Action classification or Action recognition is the ability of intelligent observers of matching the perceived actions of objects with the memorized classes of actions. The classes of actions are finite mental representations describing and grouping the infinite possible actions. The classes of actions and their defining proprieties are stored in the long-term memory of the observers.

In some studies (Chaaraoui et al., 2012; K. Lee et al., 2015; Liu et al., 2015; Y. Liu et al., 2016), they distinguish activities and actions. Generally, they define activity as a complex sequence of simple actions. For instance, Chaaraoui et al. (2012) claim a hierarchical structure of the human behaviour. At the bottom of this hierarchy, there are motions which can be detected in a time frame in units of ms. At the higher level there are the actions which are sets of motions and have a time frame in units of seconds. Activities are at the upper level with a time frame in units of minutes and are sets of actions. At the top level, there are behaviours which can be detected in a time frame in units of days or weeks. Behaviours are sets of activities and describe habits and routines. Despite of the differentiation of activities and actions made by some studies, I am going to use the words action and activity as synonymous in this thesis.

What is actually an action? Herath and colleagues (Herath et al., 2017) claim that an "action is the most elementary human-surrounding interaction with a meaning". The interaction is the relative movements with respect to the surrounding objects which may or may not change. The meaning of the interaction defines the category or class of action. However, they do not specify how the meaning or the class of an action to objectively assess whether an action fall in specific action class such as brushing hair or chopping onions. Wang and colleagues (Wang et al., 2016) clearly specify what define each meaning or class of actions. The meanings of the actions depend on what the action causes in the environment. They argue that "the true meaning of an action lies in the change or transformation an action brings to the environment" and the action "changes the state of the environment from what it was before the action to what it will be after it". Similarly, Wurm and Caramazza (2022) suggest that to study action recognition at the conceptual level, we need to focus on the action aspects that capture the what the actions actually cause (the effects of the actions) and exclude related actions aspects such as the how the actions are done (the specific movements of the actions; with hand or foot?) and the why of the actions are done (motivation, intentions, goals). Intentions and goals are the outcomes that effectors aim and must not be confused with the actions. Intentions and goas generally match the actions, but not always: a person can have the goal to kick the ball and fall down instead.

My definition of action is closer to Wang et al. (2016) and Wurm and Caramazza (2022). I define actions or activities of some actors as changes of states of recipients made by some serial or simultaneous movements of the effectors' body parts with some instruments (or tools) regardless of the specific movements of the actors. In the whole thesis, actors and agents are synonyms and I will refer to them as the animated and inanimate objects (effectors, recipients and instruments) that participate in the actions.

The changes in recipients' states happen over time and, therefore, they are definable by velocity and acceleration from some states to others. The velocity of a state change is the first derivative of the recipients' states over time and its acceleration is the second derivative of the recipients' states over time. An instance of an action is cutting/chopping which is definable as the change of state from larger and fewer pieces to smaller and more pieces made by a man (effector) to some onions (recipients) with his right hand (tool_1) and a knife (tool_2) on a cutting board (tool_3) placed on a table (tool_4). Furthermore, the action is independent of the specific moments involved in the action. The previous example of action is independent of whether the action was performed with either the left or right hand, with either knife or other sharp tools on either the table or on the chair, etcetera. The action (chopping) is only defined by the change (from larger and fewer pieces to smaller and more pieces) made by the effectors' body parts (man) to the recipients (onions) with some instruments (right hand, knife, cutting board, table).

The total possible actions within each action class are infinite by definition because the actions within the same action class, like cutting/chopping, are defined by continuous variables such as pieces size, time, velocity and acceleration of the change. A continuous variable (for instance, velocity) is defined by having infinite values within the range of any two different values (for instance, 3.0 cm/s and 3.001 cm/s) regardless of how tiny that range is (0.001 cm/s). Therefore, the possible actions within the same class of actions are infinite because they can be performed with infinite combinations of infinite values of sizes, times, velocities and accelerations and so on. Yet, the observers classify these infinite possible actions into the same action class that is cutting/chopping.

1.2.1 Bayesian Inferences in Object Classification

Let me assume the same images of the video v which I defined in Equations 1.1 and 1.2. However, the observer's task is now estimating the j^{th} action class a_j in the video v. Then, for A possible action classes where A is a positive integer, the posterior probability $P(a_j|s_t)$ of the j^{th} action class a_j given the image s_t is defined as:

$$P(a_j|s_t) = \frac{P(s_t|a_j) \times P(a_j)_t}{P(s_t)} \qquad 1.6$$

j is any integer in the range 0 ≤ j < A. $P(s_t|a_j)$ is the conditional probability of the image s_t given the action class a_j. $P(a_j)_t$ is the prior probability of the action class a_j at the timepoint t regardless of the observed image s_t. As shown in Equation 1.7, $P(a_j)_t$ is equal to the prior probability $P(a_j)$ for the first timepoint with t=0. $P(a_j)$ is independent of the image s_t. For t>0, the prior probability $P(a_j)_t$ at the timepoint t is updated to the previous posterior probability $P(a_j|s_{t-1})$ of the action class a_j given the previous observation s_{t-1}.

$$P(a_j)_t = \begin{cases} P(a_j), & if\ t = 0 \\ P(a_j|s_{t-1}), & else\ if\ t > 0 \end{cases} \qquad 1.7$$

Finally, the observer evidence $P(s_t)$ of the image s_t in Equation 1.6 is defined as:

$$P(s_t) = \sum_{k=0}^{A-1} P(s_t|a_k) \times P(a_k)_t \qquad 1.8$$

where $P(a_k)_t$ can be defined by replacing all j with k in Equation 1.7.

1.3 Advantages of Classifying Objects and Actions

By classifying both infinite objects at infinite possible states and their infinite possible actions into finite O object classes and finite A action classes, the intelligent observers solve two major issues. One, classifying actions and objects makes the unknown actions and the unknown objects known (Moscovici, 2001; Voci, 2003) such that they can react appropriately to them. Given that there are infinite possible actions and infinite possible objects in the real world defined by continuous variables, observers very often see unknown actions and unknown because their experience is finite and could not perceive and memorize infinite possible actions and infinite possible objects at infinite possible states. By classifying these unknown actions into some known action classes with known actions features and these unknown objects in some known object classes with known objects features, they retrieve the known features of these action classes and the known features of these object classes from their long-term memory. Then, they use these features of action classes and feature of objects classes to describe and embody the perceived unknown actions and unknown objects. This last step makes the unknown actions and the unknown objects known.

Two, they significantly reduce the computational cost in perceiving, memorizing, representing the infinite actions and infinite objects of the world. By definition, they cannot perceive and memorise infinite action and object information with limited neurological resources. Thus, instead of dealing with infinite actions and infinite objects, they only need to perceive and memorize finite A possible action classes with the reliable action features of each action class and finite O objects classes with the reliable object features of every object class. In this way, they save computational cost because they can neglect and forget all differences of action features within action classes and all differences of object features within object classes. For instance, after classifying 8 lines into 2 classes of 4 lines, humans underestimate the within-group differences of the line lengths and overestimate the between-group length differences of the lines (Tajfel & Wilkes, 1963)

There are at least four other reasons why observers classify actions and objects. One, action classification can ease object classification (Blasing & Sauzet, 2018; Mitchell & Curry, 2016). Mitchell and Curry (2016) showed that human observers accurately recognised some known human actors from their walks which were presented as point-light displays. Point-light displays are dark videos showing only some moving lights attached to the main body joints (ankles, knees, wrists, elbows, shoulders and more) of some actors while these actors are performing some actions. The participants could not see the actors in their point-light displays, and they could only see their walking movements. Therefore, they accurately recognised the actors from their point-light displays because they recognised the action movements from the point-light displays and then they recognised the actors from the recognized actions. This effect can be explained by revising the estimations of the Bayesian observers that classify objects given some observations. These were described in sub-section 1.1.1. By classifying the action into a class of actions a_j, observers can utilize the posterior probability $P\left(o_i \middle| a_j\right)$ of the ith object class o_i given the action class a_j as prior probability $P(o_i)_{t=0}$ of the object class o_i at first timepoint 0 to efficiently estimate posterior probability $P(o_i|s_{t=0})$ of the ith object class that the actor belongs to given the first image $s_{t=0}$ of the video v.

Two, object recognition facilitates action recognition (Ferstl et al., 2017; Knoblich & Sebanz, 2006; Schutz-Bosbach et al., 2006; Sebanz et al., 2006). Observers recognise the action of known objects with known probability distribution of the action classes given some observations more accurately and more quickly than the action of unknown objects with unknown probability distribution of possible action classes. Assuming that the observers are Bayesian estimators of the hidden states of the world given their observations, then the observers can use the known prior probability $P(a_j|o_i)$ of action class a_j given a known object class o_i as prior probability $P(a_j)_{t=0}$ at timepoint 0, to estimate the posterior probabilities $P(a_j|s_{t=0})$ of the action class a_j of that object given some observation $s_{t=0}$ at the first observation. In this way, their estimations converge more accurately and more quickly with this more informative prior probability.

Three, by classifying actions and objects, observers learn the prior probabilities of action classes and object classes. For instance, observers would predict nearly zero probability that some humans fly because they have never seen flying humans before. However, if they started seeing some flying humans, they would learn that the humans can fly sometimes. Following that, they would predict a higher probability of flying given some humans. In Bayesian terms, by classifying the new observed actions and actors into the jth action class a_j and the ith object class o_i, they learn the prior probability $P(a_j)$ of the jth action class a_j and the posterior probability $P(a_j|o_i)$ of action class a_j given the object class o_i. Observers will then utilize this knowledge to define the prior probability $P(a_j)_{t=0}$ and to calculate the posterior probability $P(a_j|s_{t=0})$ of the action class a_j given the first observation $s_{t=0}$ at timepoint 0. At the same time, they learn the prior probability $P(o_i)$ of the ith object class o_i and the posterior probability $P(o_i|a_j)$ of object class o_i given the action class a_j. Observers later use either of them as prior probability $P(o_i)_{t=0}$ at timepoint 0 to estimate the posterior probability $P(o_i|s_{t=0})$ of the ith object class o_i given the first observation $s_{t=0}$.

Four, they learn the personality traits and mental states such as intentions, feelings (Foulsham & Lock, 2015) and thoughts of the actors by the classes of actions they tend to do. Let me define x as a list of M personality traits and mental states where M is the number of all traits and mental states. By knowing the personality traits and mental states x of the actor, observers can also use the conditional probability $P(a_j|x)$ of the action class a_j given x as prior probability $P(a_j)_{t=0}$ of the action class a_j at the first observation $s_{t=0}$ and classify the actions of that actor in some images more accurately and more quickly. The reason is that their estimation of the posterior probabilities $P(a_j|s_t)$ of the action class a_j given some observation s_t in Equation 1.6 converges quicker with a more informative and more specific prior such as $P(a_j|x)$.

1.4 Three Visual Pathways Process the Subskills of Active Action Classification

According to three-visual-pathway theory (Boussaoud et al., 1990; Galletti et al., 2003; Murata et al., 2016; Rizzolatti & Matelli, 2003; Tanne-Gariepy et al., 2002; Wurm & Caramazza, 2022), there are three visual pathways or streams of hierarchical cortical areas that share the lower levels areas and dissociate in higher-level areas. The three visual streams are the ventral, the lateral and the dorsal streams. The ventral stream includes the occipital cortex and inferior temporal cortex (IT). The dorsal stream contains the occipital cortex, superior parietal cortex (SP) and superior (or dorsal) premotor cortex (dPM). The lateral stream comprises the occipital cortex, the middle and superior temporal cortex (MT and ST), inferior parietal cortex (IP) and the ventral premotor cortex (vPM). The different visual subskills of the active action classification seem to rely on three different visual pathways. Object classification is a task of the ventral stream. The online egocentric localization for navigation control is processed by the dorsal stream. However, the lateral stream seems to be involved in several tasks related to social cognition which include action classification (Wurm & Caramazza, 2022), visual perspective taking (Santiesteban et al., 2015; Schurz et al., 2013) and even theory of mind (ToM; attribution of mental states to others) (Santiesteban et al., 2015).

All three visual pathways are hierarchical because their areas process the stimulus features with different levels of the abstractness. The lower-level visual areas process more concrete visible features such as edges, edge orientations and colours, while the higher-level areas embody more abstract features of the world such as object classes, actions classes and locations. That is in line with the fact that retinotopy in higher-level areas decreases (Malach et al., 2002) while receptive field and supramodality increases. In fact, the neurons of the lower-level neurons tend to have smaller receptive fields and a more retinotopic organization, whereas the higher-level areas have wider receptive field and are less retinotopic. Additionally, lower-level visual areas are less supramodal such as retinal ganglion cell (RGC), lateral geniculate nucleus (LGN), V1, V2, V3, V4, V5 and V6 that only respond to visual stimuli, while the higher-level areas are more supramodal.

High level areas that embody abstract concepts such as object classes and actions classes are expected to be supramodal because they should generalise across stimulus modalities and they should be independent from each stimulus modality. For instance, the V6A and SP and IP areas have neurons that respond to somatosensory and visual signals. The IP and vPM areas contain mirror neurons that are sensitive to the observation and execution of actions. The vPM areas also include audiovisual mirror neurons that respond to the observation, hearing and motor execution of actions. The middle temporal gyrus (MTG) and superior temporal sulcus (STS) areas respond to observation or hearing of actions or simple sentences describing actions.

The three hierarchical visual pathways share a common origin of lower-level visual areas and dissociate at their higher-level areas. They mostly share the visual areas in the occipital cortex and segregate in the temporal, parietal cortex. Overall, the ventral stream includes the occipital cortex and IT. The lateral steam includes occipital cortex ST, IP and vPM. The occipital cortex, SP and dPM cortex belong to the dorsal stream. The lower-level visual areas that the three pathways share are the RGC, LGN, V1, V2. Next, the ventral stream continues with V4 and ends with IT. Nonetheless, the areas V3 and V3A do not belong to the ventral stream, they are part of and shared by the lateral and dorsal streams. After, the lateral streams continue with MT (or V5) which projects to the ST including medial superior temporal area (MST) and to IP. Finally, the lateral stream end at the areas F4 and F5 in the vPM which manly control the ventral primary cortex that moves mouth, tongue, lips and face. The area F5 is the Brocka's area. The IP and the vPM contains mirror neurons which are known to be involved in action understand. In the dorsal pathway, after V3 and V3A, there are V6 and V6A areas which both make the parietooccipital area (PO) in the parietooccipital sulcus (POs). Following V6 and V6A, there are some areas on SP which project to the dPM which directly controls the activity of the dorsal primary motor cortex.

It is unclear whether each visual area in a visual pathway is equivalent to a layer in NNs. On one hand, Liao and Poggio (2016) argued that if we assume that each area in the ventral pathway corresponded to a layer in NNs, then the ventral pathway would be a shallow recurrent neural network (RNN). It would be shallow because there are about six visual areas in the ventral pathway, while modern ultra-deep NNs have hundreds of layers (He, Zhang, Ren, & Sun, 2016; Szegedy et al., 2015). It is a RNN because it has complex temporal loops of the information flow via lateral and backward (feedback) connections. Liao and Poggio (2016) suggested that the visual ventral pathway is a multi-stage processing hierarchy with full recurrent connections. Each area of the pathway receives inputs from all lower areas by forward connection, from all higher areas by feedback connections and from itself by lateral connections. Therefore, there are forward, lateral and backward connections in the visual steams (Kar et al., 2019; Kubilius et al., 2018; Lamme et al., 1998). Furthermore, the forward and backward connections are simple or shortcut connections. Shortcut connections (He, Zhang, Ren, & Sun, 2016; C. Y. Lee et al., 2015; Schraudolph, 1998; Srivastava et al., 2015; Szegedy et al., 2015) enable the neural activation of a layer to skip one or more subsequent layers. The backward connections inspired top-down theories such as predictive coding (Rao & Ballard, 1999) and free energy principle (Friston, 2003; Friston et al., 2006). Regardless of these interpretations about their backward connections, these areas are technically RNNs because the activations of the areas in a timepoint are influenced by their own activations in the previous timepoint.

Liao and Poggio (2016) added that if the ventral pathway were a shallow RNN, then it would be very efficient and it is supposed to be efficient because of evolution. It would have two main advantages if it were a shallow RNN. The first advantage would be that its depth (number of layers) would be flexible given that it could be both shallow with short processing time and ultra-deep with long processing time (Liao & Poggio, 2016). By unrolling the information flow in time, the depth of a RNN is equal to T which is the total number of the neural activation timepoints and is a positive integer in the range $0 \leq T \leq +\infty$. In fact, the responding activity of a shallow RNN goes through T layers or through the same layer for T times. Thus, a RNN can be shallow with a few timepoints and can also be ultra-deep with many timepoints up to positive infinity. We can assume that the duration of a neural activation timepoint can approximately be 20-50 ms like some evidence suggest. For instance, there are 100 timepoints (T=100) for a neural processing time of 2 seconds if we set the duration of a neural activation timepoints to 20 ms. Anyway, because the depth of the brain areas would be flexible and would depend on the neural processing time, then the ventral pathway could process faster and less accurate responses with only a few activation timepoints. It could also process long and more accurate responses with abundant activation timepoints. The second advantage would be that it would keep the number of neurons (model parameters) low while being ultra-deep in time because RNNs share neurons (parameters) across time, contrarily to simple forward NNs (Liao & Poggio, 2016).

However, Storrs and colleagues (2021) fitted the activation of layers of common convolutional neural networks (CNNs) to the human IT activation during the observation of objects in images. They found no relation between the number of layers of the models and the degrees of models' fitting to the IT. The layers of the models with hundreds of layers fitted the IT as well as the ones of the models with a few layers. Therefore, their findings do not support the hypotheses of Liao and Poggio (2016) by which each area of the ventral system corresponds to a NN layer and the ventral system is a shallow RNN.

1.4.1 Historic Development of the Three-Visual-Pathway Theory

The 3-visual-pathways theory has roots in Ungerleider and Mishkin (Mishkin et al., 1983; Ungerleider & Mishkin, 1982). They originally claimed there were two visual pathways: the ventral and dorsal pathways. Anatomically, the ventral stream includes the occipital areas and IT whereas the dorsal stream consists of the occipital and the parietal cortex. According to them, the ventral stream functions as the object perception and object discrimination whereas the dorsal stream as perception of space and location of objects. In fewer words, the ventral stream perceives the “what” while the dorsal stream perceives the “where” of the objects.

Boussaoud and collegues (Boussaoud et al., 1990) were the first neuroscientists suggesting there visual streams. For them, the ventral and the dorsal streams are anatomically functionally and are identical to original theory of Ungerleider and Mishkin (Mishkin et al., 1983; Ungerleider & Mishkin, 1982). Nevertheless, the anatomy of the third visual pathway contains the occipital cortex and ST, while its functions is motion analysis. This motion analysis supports both the ventral stream for the object perception and the dorsal stream for the spatial perception, although more to the dorsal than to the ventral.

Ten years later, Goodale and Milner (Goodale & Milner, 1992; Milner & Goodale, 2008) agreed with Ungerleider and Mishkin (Mishkin et al., 1983; Ungerleider & Mishkin, 1982) about the anatomical and functional distinction of the two visual pathways. However, they disagreed with the actual functions the two streams do. They said that “both steams process information about the structure of the objects and about their spatial locations”. But the streams process visual information in different ways for different visual skills. The ventral stream processes the vision for perception and dorsal stream process the vision for action. On one hand, the ventral stream produces conscious perceptual representation of the characteristics of the objects. On the other hand, the dorsal stream processes the visual information to control object-directed movements of actions, such as picking up a mug and pouring water into a glass) which are directed to target objects. To control object-directed movements, the dorsal stream mainly processes and update the egocentric coordinates of the target object on the moment-to-moment basis because these egocentric coordinates quickly change as movements are being executed. Because the dorsal stream estimates the egocentric coordinate of the target objects in real-time, the dorsal stream relies more on the bottom-up visual information than the ventral stream. Therefore, for Goodale and Milner, the ventral stream processes “what” the objects are, and the dorsal stream processes the “how” to manipulate them.

Some other researchers (Galletti et al., 2003; Murata et al., 2016; Rizzolatti & Matelli, 2003; Tanne-Gariepy et al., 2002) have anatomically describe the three visual pathways by dividing the dorsal stream in two: the ventro-dorsal (lateral) and dorso-dorsal streams. They all agree with Goodale and Milner (Goodale & Milner, 1992; Milner & Goodale, 2008) about the function of the ventral stream. The function of the dorso-dorsal stream corresponds to dorsal stream of Goodale and Milner. They claim different functions to the ventro-dorsal stream. For example, Rizzolatti and Matelli (2003) said that the ventro-dorsal stream encodes space and action understanding because the ventro-dorsal stream is rich of mirror neurons.

Wurm and Caramazza (2022) proposed a new model with three visual pathways. They agree with Boussaoud and collegues (Boussaoud et al., 1990) about the anatomy of the three pathways. However, Wurm and Caramazza diasgree with Boussaoud and collegues about the functions of the three pathways. Wurm and Caramazza agree with Goodale and Milner (Goodale & Milner, 1992; Milner & Goodale, 2008) about the functions of the ventral and dorsal streams. The third lateral pathway encode for abstract conceptual social actions.

1.4.2 Lower-Level Visual Areas

RGC seems to process a retinotopic map of edges of the visual stimuli. Kuffler (1952) noticed that RGC of a cat are specifically sensitive to edges placed on their corresponding receptive fields. He modelled the activity and the receptive fields of the RGC by DoG. Then, DoG was then used in computer vision for edge detection (Basu, 2002; Kennedy & Basu, 1997; Marr & Hildreth, 1980; Wohrer & Kornprobst, 2009)

The area V1 contains several retinotopic maps for all edge orientations of the visual stimuli: one retinotopic map for each edge orientation. Hubel and Wiesel (1962) recorded the activity of the neurons in V1 of a cat while stimulated visual stimuli. They noticed that the activities of these neurons were sensitive to specific edge orientations positioned on their receptive field. Likewise, Marcelja (1980) suggested that Gabor filters can model the receptive field of either LGN or V1. The Gabor function was originally presented by the Hungarian-British physicist Dennis Gabor (1946) and was then extended to 2d filters by Daugman (1985). The Gabor filters have been good tools in visual tasks. For instance, Rizvi et al. (2016) showed that a NN aided with a bank of Gabor filters in the first layer got comparable accuracy (50.71) to the baseline convolutional neural network (52.15) in object recognition. See Rai and Rivas (2020) for literature review about Gabor filters.

1.4.3 The Highest-Level areas for Object Classification are in IT

What are the cortical areas that embody the object classes at their highest-levels? Despite of some conflicting results, several functional MRI (fMRI) studies point their fingers to IT as the best candidate for this role. To be as such, the highest-level areas of object classes must have wide receptive fields, not be retinotopic, and be supramodal. Adams and Janata (2002) did a fMRI study revealed that neural circuits underlying auditory and visual object categorization share IT and the middle and inferior frontal cortex. Fairhall and Caramazza (2013) asked to participants to classify objects that could either be presented as written texts or as pictures inside a fMRI scanner. Both tasks activated the middle and inferior temporal cortex and posterior cingulate.

Man et al. (2015) made a fMRI study with either visual, tactile or even auditory object recognition tasks. They also found an overlapping activation in IT during the visual and tactile object recognition, and even in posterior ST, IP, SP and lateral and ventral occipital cortex. However, there were only overlaps in the posterior ST, IP and lateral and ventral occipital cortex and there were no overlaps in IT. The activities lateral and ventral occipital cortex should have been some low-level visual representations of the objects which may triggered in the visual condition through bottom-up feedforward connections and top-down feedback connections. These may have triggered even in the tactile and auditory conditions by top-down feedback connections. The posterior ST and IP may contain more supramodal, abstract and semantic representations describing actions, relations and mental states of objects. These may have been activated by either the observation, the hearing or the touch of either the effectors, recipients or tools which are associated to the actions even without seeing, hearing or touching the actions themself. The participants actually saw the actions in the visual and auditory conditions because the visual and auditory stimuli were videos and sound of objects in action. However, the action representations may have been activated only by the touch of the object and without any perception of the actions in the tactile condition. There should have been activity overlaps on IT in all visual, tactile and auditory conditions, if IT really embodies abstract aspects of the object classes which are related to their appearance and not related to their potential actions and mental states. However, it is unclear why there was only overlaps in IT in the visual and tactile conditions while IT activity did not vary in the auditory condition.

Some (Ishai et al., 2000; O'Craven & Kanwisher, 2000) may argue that this overlap of neural activity may in IT may be the result of a visual imagination that may be even triggered by either the touch, the hearing of the objects or the observation of their written names rather than being supramodal object representations. Nonetheless, Pietrini et al. (2004) took fMRI scans of sighted participants while they were recognising objects by either the vision of the objects in images or the touch of these. The found an overlap of brain activity in IT and ventral temporal cortex during both visual and tactile object recognition. Importantly, they also took fMRI scans of blind participants while doing the same tactile object recognition and found a similar activation of IT. The blind subjects were either congenitally blind or had become blind at an early age and reported no visual memories. Therefore, the hypothesis that IT process visual imagination during tactile object recognition should be excluded and IT may really encode abstract and supramodal representation of objects classes.

It is interesting to note that Amalric and Dehaene (2016, 2018, 2019) highlighted IT and intraparietal sulcus seem to be also involved in both basic and high-level mathematic and geometric calculations. Given that IT is involved in object recognition, some may argue that the IT activation during mathematics and geometry may have processed object recognition of the visual numbers, words and letters in the stimuli rather than processing mathematic and geometric computation. However, all mathematical and geometric questions (stimuli) were auditory (spoken to the participants) in their studies rather than visual. Thus, it very likely that IT embody mathematic and geometric entities which can be thought as abstract, supramodal and high-level features of the object appearances.

1.4.4 The Highest-Level Areas for Action Classification are in MT and ST

Action recognition is associated with higher activity of regions in the frontal, parietal and lateral temporal cortex (Buccino et al., 2004; Hamzei et al., 2003; Lingnau & Downing, 2015; Oosterhof et al., 2013; Orban et al., 2021; Wurm & Caramazza, 2022). The posterior areas, like temporal and parietal cortex, are anatomically closer to the low-level visual areas, like V1 and V2, rather than the more anterior areas, like the frontal cortex. Therefore, an intuitive interpretation is that the more posterior areas in temporal and parietal cortex process less abstract aspects of the actions, like objects, postures, and movements, whereas the anterior areas in frontal cortex process the most abstract action features, like action classes. The findings in the literature seem to be against this intuitive view.

The highest-level areas for action classification encode the most abstract action classes. I formulate two necessary criteria to identify these areas. One, they must encode supramodal representations of actions. This means that they must similarly respond to the same actions, even if these actions are presented in different sensory modalities or formats. For example, they must similarly respond to the same actions regardless of whether the participants either observe the actions, hear sounds of the actions, read sentences describing the actions, or hear verbal descriptions of the actions. Two, they must encode the actions classes which are independent from the specific movements of the effectors. Therefore, they must not be sensitive to the different movements of different actions that fall into the same action category. For example, their neural activity must be similarly sensitive to different actions which involve different movements (drinking water with right hand and drinking water with the left hand) but they conceptually fall into the same actions class (drinking water).

The response of LT and IP to actions are independently from either the specific body parts (Vannuscorps et al., 2019) or movements (Wurm & Lingnau, 2015) involved in the actions. The neural response of the frontal cortex, like the PM, to actions is sensitive to the specific body parts (Hafri et al., 2017) and moments of the actions (Wurm et al., 2017).

The responses of all the frontal, parietal and lateral temporal cortical areas to actions are supramodal. In the premotor cortex (PM) and IP, there mirror neurons (Buccino et al., 2001; Buccino et al., 2004; Chong et al., 2008; Gallese et al., 1996, 2002; Hamzei et al., 2003; Rizzolatti et al., 1996) which respond to either the execution or observations of actions and audiovisual mirror neurons (Keysers et al., 2003) which respond to the execution, vision or sound of actions. The posterior ST and IP contain also semantic language areas including the Wernicke's area (Binder, 2015). Lesion on the Wernicke's area causes Wernicke's aphasia (Yang et al., 2008) which is the inability to comprehend sentences and the production of fluent and meaningless sentences. Additionally, Foxe et al. (2002) also show the supramodal nature of ST because they also found activity overlaps in the ST by either the auditory or tactile stimuli.

Wurm and Caramazza (2019) took fMRI scans of human participants while they were either watching videos of actions or reading sentences that describe the same actions. They performed both unimodal and crossmodal multivoxel pattern analysis (MVPA) (Kriegeskorte et al., 2006) to identify brain regions that encodes unimodal and supramodal action representations, respectively. In both analysis types, they trained classifiers to decode the action classes from the activities of the brain regions which were stimulated by either videos or sentences. In the unimodal (within-modality) analysis, the classifiers were trained on the brain activity in some trials of one stimulus modality and tested on the brain responses of some other trials within the same modality. In the within-video analysis, they trained the classifiers on the brain data in some video trials and tested them on the same data in some other video trials. In the within-sentence analysis, they trained the classifiers on the brain responses of some sentence trials and tested them on the brain responses of some other sentences trials. In the crossmodal (between-modality) analysis, they trained the classifiers on the neural activity of the videos and tested them on neural responses of the sentences. The action decoding accuracy of within-video and the within-sentence unimodal analyses overlapped in some cortical areas of the PM, IP and posterior lateral temporal (LT which includes MT and ST). The overlapping areas identified by unimodal analyses were in line with other studies (Aziz-Zadeh et al., 2006; Spunt & Lieberman, 2012). These findings suggest that the overlapping areas encode supramodal representations of actions. However, they were only able to decode the actions in the LT from the crossmodal analysis. Therefore, the frontal and parietal cortex encoded modality-specific action representations, whereas LT encoded the supramodal action representations of action classes.

Taking together the findings of these studies, the posterior MT and posterior ST in the posterior LT are the only cortical areas that encode supramodal representations of action classes which are independent from the movements of the actors.

1.5 Deep Learning for Visual Skills of Active Action Recognition

Active action classification involves the following four visual subskills or visual tasks: object classification, action classification, visual perspective taking and egocentric localization to guide ongoing viewpoint navigation to the target position. Deep learning models can be trained to efficiently do each of the visual tasks. In object classification, these models have been trained to predict the object classes in visual stimuli (He, Zhang, Ren, & Sun, 2016; Krizhevsky et al., 2017; Liao & Poggio, 2016; Simonyan & Zisserman, 2014; Szegedy et al., 2015). The visual inputs or visual stimuli in these visual tasks are either images or videos (which are technically a sequence of images). In action classification, they have been trained to predict the action classes that are in the visual inputs (Aghaei et al., 2021; Al-Faris, Chiverton, Ndzi, et al., 2020; Al-Faris, Chiverton, Yang, et al., 2020; Dai et al., 2020; Donahue et al., 2015; Li et al., 2018; Majd & Safabakhsh, 2020; Ng et al., 2015). In visual perspective taking, they have been trained to predict what can be seen from another viewpoint given the current viewpoint (Jayaraman & Grauman, 2018). In egocentric localization, the models were trained to simultaneously predict the coordinates and the object classes of multiple objects in real-time (Redmon et al., 2016).

In 2012, the massive success of AlexNet (Krizhevsky et al., 2017) in single-image object classification showed the promising advantage of CNNs in visual tasks. AlexNet is a non-recurrent (forward) 2d CNN and its performance of 2012 come from the single images of the dataset ImageNet (Deng et al., 2009). ImageNet has thousands of object classes in millions of images. Since the success of AlexNet in 2012, CNNs have been widely used in any visual tasks. Technically, CNN are a specific type of NNs that has one or more convolutional layers, and each convolutional layer is a bank of filters with trainable parameters by gradient descent.

The architectures CNNs can generally split in two model types or categories: non-recurrent and recurrent model. The non-recurrent models only have layers with forward connections and without any lateral or backward connections, whilst the recurrent models contain one or more layers with either lateral or backward connections.

1.5.1 Non-Recurrent Convolutional Neural Networks

There are at least three non-recurrent CNNs for visual tasks: 2d CNNs without temporal pooling layers (2d CNNs), 2d CNNs with temporal pooling layers and 3d CNNs.

1.5.1.1 2d CNNs

The 2d CNNs analyse single images individually and independently from the previous images (observations) of same video. In other words, they cannot integrate features of the images in the same video. Therefore, they are mostly used to analyse single image samples and they are not commonly used for videos. Some popular 2d CNNs are AlexNet (Krizhevsky et al., 2017), VGG (Simonyan & Zisserman, 2014), Inception (Szegedy et al.,

2015) and ResNets (He, Zhang, Ren, & Sun, 2016).

The general architecture of simple 2d CNNs has two types of layers: 2d convolutional layers and fully-connected layers. The lower or shallower layers are 2d convolutional layers to extract the local spatial features (like eyes, ears, nose, mouth, legs, arms, tails, wheels, doors, windows). Each convolutional layer is a list of 2d filters (kernels) with trainable parameters. The deeper or higher layers are fully connected layers which extract the global features (like people, dogs, cars, houses). There are only forward connections between layers of the models and no lateral or backward connections. There may be some forward shortcut connections in some 2d CNNs such as ResNets (He, Zhang, Ren, & Sun, 2016) and inceptions (Szegedy et al., 2015), but these models do not have backward shortcut connections.

Yamins and DiCarlo (2016) claimed that the 2d CNNs do similar computations of the visual pathways because of several reasons. One, by moving from lower to higher layers, the visual receptive field increases and retinotopy decreases, similarly to the visual areas of the visual cortex (Malach et al., 2002). Two, the first layer seems to naturally learn by gradient descent some traditional filters such as difference of gaussians (DoG) and Gabor wavelet filters (Karpathy et al., 2014; Krizhevsky et al., 2017; Kubilius et al., 2018) which neuroscientists have used to model the neural responses of the low-level areas to simple visual stimuli like edges and edge orientations.

Nevertheless, it is unclear whether hidden layers of deep layers in 2d CNNs do similar computations as middle-level areas in the visual pathways such as V2, V3 and V3A. Another limit of 2d CNNs is that there is no reference about response times (RTs). Hence, RT cannot be used as dependent variable and compare it in different experimental conditions. It is only possible to compare the precision of their predictions in different conditions of any experiment. This is mostly because they have layers with only forward connections and do not have any recurrent layers with either lateral or backward connections.

1.5.1.1.1 Residual Networks

He et al. (He, Zhang, Ren, & Sun, 2016) proposed residual networks (ResNets) which are ultra-deep 2d CNNs with hundreds and sometimes even thousands of layers with trainable parameters. He et al. claimed theoretically and experimentally that ResNets enjoy higher performances by increasing their depth while non-residual neural networks face the degradation problem. Most non-residual (or plain) 2d CNNs which perform very well in different visual tasks are relatively deep with a depth of 16 (Simonyan & Zisserman, 2014) and 30 layers (Ioffe & Szegedy, 2015). However, the accuracy of deeper plain 2d CNNs in object recognition get saturated and degrades rapidly.

The degradation problem is not caused by overfitting given that the training loss rises as well as the validation loss, by increasing the number of layers of the plain 2d CNNs. It is not even caused by the vanishing and exploring gradient problem (Bengio et al., 1994; Glorot & Bengio, 2010). This impeded very deep neural network to learn. This problem has been solved by normalised initialization (Glorot & Bengio, 2010; He et al., 2015; LeCun et al., 1998; Saxe et al., 2013) and normalization layers (Ioffe & Szegedy, 2015), which enabled deeper neural networks with up to 30 layers to converge. However, the degradation problem remains for ultra-deep neural network with over 30 layers.

If it is neither overfitting nor vanishing/exploding gradients, what can explain the degradation problem? He et al. further argue that the degradation problem is due to the difficulty of approximating a stack of multiple non-linear layers to a function *f* whose optimal output feature mapping y is equal to the input feature mapping y:

$$f(y) \approx y \qquad 1.9$$

where *f* is a stack of multiple non-linear layers and y is the inputs feature mapping and the optimal outputs feature mappings of the function *f*. In fact, let us assume we have a shallower neural network with N layers and a deeper one with M layers, where N and M are positive integers and M > N + 1. The deeper neural network has a stack of multiple L additional non-linear layers respect to the shallower one, where L is a positive integer and L > 1. Let us also suppose that the shallower neural network can optimally predict y give an input x. Then, the counterpart deeper neural network should also be able to optimally approximate the prediction y given x, if the L additional non-linear layers are able to output the feature mapping y given y. In fact, the first N layers of the deeper M-layer network would predict the optimal y given x, and the last L layers would predict y from y. However, since we experimentally observe the degradation problem with deeper neural networks, it may be difficult to optimize a stack of non-liner layers in predicting a feature mapping y from the same feature mapping y.

In deep learning, it is popular the hypothesis that multiple non-linear layers can asymptotically approximate any complicated function. This hypothesis is still open (Montufar et al., 2014). Let us call this hypothesis the any-function hypothesis. Theoretically speaking, if the any function hypothesis is true, then we should not expect any higher training error for deeper neural networks. The reason is that if multiple non-linear layers can asymptotically approximate any function, then it is plausible to assume that multiple L non-linear layers can asymptotically approximate the identity feature mapping y given the same feature mapping y. However, we have experimentally witnessed the degradation problem which is the effect of higher training error for deeper non-residual networks.

In ordinary non-residual neural network, we optimize each block of non-linear layers *f* to fit the desired feature mappings y given the input feature mappings x:

$$f(x) \approx y \tag{1.10}$$

This type of architectures faces the degradation problem. Contrarily, He at al. introduced the ResNets to tackle this issue. The ResNets are divided into numerous blocks of multiple non-linear layers which are optimised to approximate the residual mapping z.

$$f(x) \approx z = y - x \tag{1.11}$$

and then the original input mappings x are added to the residual mappings z and get the desired output mapping y:

$$f(x) + x \approx y \tag{1.12}$$

ResNets solve the degradation problem and can gain accuracy in object classification with deeper models up to hundreds of layers. According to He et al., the main reason is that the residual feature mappings z are easier to approximate than the desired mapping feature y by a block of multiple non-linear layers. Ideally, if the first blocks of non-linear layers optimally predict the desired mappings y, then it is easier to approximate the residual mappings z with any additional blocks of non-linear layers (rather than the identity feature mappings y), by pushing all their weights close to zero.

Let us define a ResNet building block of non-linear layers. For simplicity, I am going to only elaborate the building blocks of fully-connected layers. However, the same principles apply to convolutional layers. There are types of building blocks in ResNets. A building block can be with either an identity shortcut or a projection shortcut. A ResNet building block with identity shortcut is defined as:

$$y = f(x, W_i) + x \tag{1.13}$$

where the size of the input mappings x and the output mappings y are equal. The function *f*(x, W_i) is a block of non-linear layers with weights W_i, where i is the number of layers within the building block. If i = 2, then *f*(x, W_i) = W_2 σ(W_1 x) where σ is ReLU (Nair &

Hinton, 2010), a non-linear activation function defined as σ(x)= max(0, x). During the training, the weights W_i are optimized such that the function *f*(x) approximates the residual mappings z. The operation *f*(x) + x is element-wise addition and is the identity shortcut. In a ResNet block, the shortcut skip i layers. In a block with identity shortcut, the residual mappings *f*(x) and the input mappings x have the same size. The identity shortcut is a simple element-wise addition which do add extra parameters to the overall model. An identity shortcut keeps the identity mappings x. After the addition, A ReLU activation function is applied to y.

In a bock with projection shortcut where, the shortcut is performed by the element-wise addition between the residual mappings *f*(x) and a linear projection of input mappings x. The liner projection requires some additional trainable weights W_s. Therefore, the formal equation of the building block becomes:

$$y = f(x, W_i) + W_s x \tag{1.14}$$

The sizes of the residual mappings *f*(x) and the input mappings x can either be different or the same. The authors mostly used projection shortcuts only in blocks where the residual mappings *f*(x) and the input mappings x are different in size, to reduce the number of parameters of the models. In this way, they mostly project input mappings x to a space of the same dimensionality as the residual mappings *f*(x), only when the sizes of *f*(x) and x are different. However, projection shortcuts can also be used for blocks where *f*(x) and x have the same size.

They showed ResNets enjoy higher accuracy by increasing the number of layers to up 152 layers. This was valid for ResNets with either only identity shortcut, only projection shortcuts or both. The identity shortcuts only were enough to solve the degradation problem without adding any extra parameters. According to He et al., the reason why deeper ResNets perform better is because approximating the residual mappings z with non-linear layers is easier than the identity mappings y. However, they have reported slightly higher performance for ResNets with only projection shortcuts than the ResNets with only identity shortcuts. This suggests that the explanation of the He et al. about the better performance of deeper ResNets may not be the only one.

According to Liao and Poggio (2016), the ResNets are very accurate because they approximate RNNs and RNNs generalize the data better than their corresponding non-recurrent neural networks (NNs) because of several reasons. First, by assuming that the biological brain is an efficient information processing system due to the evolution of the species and most visual and non-visual areas of the biological brain are RNNs, then RNNs must be more efficient than simple forward NNs. Second, RNNs generally outperform their corresponding non-recurrent NNs in several visual and non-visual tasks, in spite of having less parameters than their corresponding non-recurrent NNs (Du et al., 2019). Therefore, an apparent ultra-deep forward ResNet can outperforms most of other forward NN, if ResNets can indeed approximate RNNs.

Liao and Poggio (2016) argued that ResNets are time-variant RNNs because they do not share parameters in time. The building blocks of an ultra-deep forward ResNet which do not share parameters in time can be approximated by less recurrent ResNet building blocks which share parameters in time. By unfolding the information flow of the recurrent ResNet building block, we get back to an ultra-deep forward ResNet with shared parameters among all blocks and its depth is the number of timepoints T where $1 < T < +\infty$. Like any RNNs, ResNets with shared parameters are therefore deep in time, while they keep the number of parameters significantly low. They experimentally showed that recurrent ResNets with shared parameters approximate the performance of their corresponding standard forward ResNet with non-shared parameters on the CIFAR-10

(Krizhevsky & Hinton, 2009) and ImageNet (Deng et al., 2009; Russakovsky et al., 2015) datasets. This result is in line with their claim that ResNets are in fact RNNs.

Liao and Poggio (2016) also argued that ultra-deep ResNets and the visual ventral stream of the brain are not very different in depth as they appear to be. On one hand, ultra-deep ResNets approximate shallow RNNs which can be very deep in time. The visual pathways like the ventral stream are shallow RNNs with forward, lateral, backward connections which are also ultra-deep in time. Therefore, the visual streams can also be ultra-deep in time as much as any ultra-deep ResNet and are not actually different in depth.

1.5.1.2 2d CNNs With Temporal Pooling Layers and 3d CNNs

Some studies (Karpathy et al., 2014) have used CNNs on videos by simply adding some temporal pooling layers that integrate the features of images of the same video. Anyway, CNNs with the temporal pooling layers gain very little performance than the CNNs on single image. There are generally better model types for videos. Some other studies (Ji et al., 2013; Tran et al., 2015) have used 3d CNNs to analyse videos. 3d CNNs have 3d (instead of 2d) kernels to extract spatiotemporal features of the videos.

However, it is hard to believe that human brain temporally integrates observations by either temporal pooling layers or 3d CNNs. Additionally, the video length (number of images), that either 2d CNNs with temporal pooling layers or 3d CNNs can be fed, is fixed (not flexible) and cannot be changed without adjusting the architecture of the models and retrain them.

1.5.2 Recurrent Convolutional Neural Networks

Recent studies (Donahue et al., 2015; Kubilius et al., 2018; Li et al., 2018; Liao & Poggio, 2016; Ng et al., 2015) have suggested recurrent convolutional neural networks (RCNNs) for visual tasks. RCNNs are a specific type of CNNs which have one or more recurrent layer with either lateral or backward connections. The recurrent layers can either be 2d recurrent convolutional layers or 1d recurrent fully-connected layers. Roboticists (Donahue et al., 2015; Ng et al., 2015) originally designed RCNNs for videos analysis because they can integrate neural activations across time (across video images) by lateral (and rarely backward) connections. However, they can potentially be used to analyse single images by feeding the models the same image at different timepoints (Liao & Poggio, 2016).

On one hand, roboticists designed RCNNs that achieve the best performance in a specific visual task, even if they are not biological plausible. On the other hand, computational neuroscientists have made RCNNs that are more biological plausible and simulate behaviour performance of humans such as accuracy and RTs in a variety of visual tasks, even if they are not the models with the best performance in a specific visual task.

Anyhow, the RCNNs of both roboticists and computational neuroscientist have many advantages compared to the non-recurrent CNNs. One, similarly to both the non-recurrent 2d CNNs with temporal pooling layers and the 3d CNNs, RCNNs can extract spatiotemporal features and make robust predictions with these spatiotemporal features at each timepoint because their predictions take into account all seen observations (video images) from the first image to the current one. The prediction of a video by a RCNNs at timepoint t, where $0 \leq t < +\infty$ is an integer, is technically a function of all images of the video at the timepoints from 0 to t.

Two, contrarily to both the non-recurrent 2d CNNs with temporal pooling layers and the 3d CNNs, the video time length (number of video images) is flexible with RCNNs (Donahue et al., 2015). The same RCNN can make the predictions from the spatiotemporal features of videos with different time lengths ranging from 0 to $+\infty$ images without changing their architecture and retraining them.

Three, they are more biological plausible because the any (visual and non-visual) cortical area of the brain are technically recurrent with forward, lateral and feedback connections (Kar et al., 2019; Kubilius et al., 2018; Lamme et al., 1998).

Four, the depth of recurrent models is flexible (Donahue et al., 2015; Liao & Poggio, 2016) because they are deep in time. By unfolding in time the information flow of RCNNs, its corresponding non-recurrent CNNs has layers that share trainable parameters and has depth equal to T where $0 \leq T \leq +\infty$ is the number of stimulating timepoints. Their neural activation goes through the same layers T times. Therefore, they can be both shallow with a few discrete timepoints and ultra-deep up to positive infinity with numerous timepoints. Therefore, they can both do fast (short) and slightly accurate predictions, and slow (long) and very accurate predictions depending on how much time is available for the responses.

Five, RCNNs maintain small their total number of trainable parameters (Donahue et al., 2015; Liao & Poggio, 2016; Ng et al., 2015), while they can still be ultra-deep. This saves a lot of computer memory. By unfolding in time a RCNN with T timepoints, it corresponds to a non-recurrent CNN with same depth T and with layers that share trainable parameters. If T is an ultra-large integer, then a RCNN is ultra-deep and has less trainable parameters than an its corresponding non-recurrent CNNs with layers that do not share trainable parameters.

Six, they can theoretically simulate different RTs in different experimental tasks that which are assumed to have different levels of difficulty (Kubilius et al., 2018; Liao & Poggio, 2016). The depth of RCNNs is the number of the activation timepoints which is flexible. Therefore, given that the number of the activation timepoints may vary across different conditions until the prediction of model gets to predefined threshold, it is possible to estimate different RTs for all conditions based on the number of the activation timepoints needed by the models to reach the prediction threshold. It is important to distinguish here the timepoints of the stimulus (videos frames) and the timepoints of the activation of the model layers. As a matter of a fact, the refresh rate of brain neurons is very often different from the stimulus refresh rate. Each stimulus timepoint is the time between a frame onset and the onset of the next one, whilst the activation timepoint of either a computational model or the brain is the time by which the activation of a layer influences the activation of the next layer. The duration of each activation timepoint of the recurrent models can be set between 20 ms and 50 ms, considering the latency of a single layer of biological neurons (Liao & Poggio, 2016). This makes an activation refresh rate between 50 Hz and 20 Hz. The RT of a video by a model can be estimated by multiplying the duration of an activation timepoint and the numbers of activation timepoints needed for the model classification to get some threshold. For instance, if we define the duration of the model timepoint to 20 ms and the classification of a recurrent model takes 100 timepoints to reach a threshold, then we would have a RT of 2,000 ms (20 ms x 100). The depth of the non-recurrent CNNs is constant and cannot not be changed once the whole model is defined. Thus, the depth of non-recurrent CNNs cannot vary across different conditions and it is not possible to estimate the RTs for different conditions. Simulating RTs is out of the scope of this thesis, but it will be the aim of my future studies.

1.5.2.1 Recurrent Convolutional Neural Networks of Roboticists

Generally speaking, an architecture of most RCNNs in robotic vision has a 2d CNN at their bottoms and the one or more 1d recurrent layers at the tops (Donahue et al., 2015; Ng et al., 2015). The 2d CNN extracts the spatial features of each observation (video image) individually and independently from the previous observations (images of the same video). The 2d CNN is often called feature extractors. The 1d recurrent layers then feed on the spatial features extracted by 2d CNN. Thus, the 1d recurrent layers extract the spatiotemporal features of the video image at timepoint t dependently from the previous images of the same video. The most popular recurrent layer is long-short term memory (LSTM) (Hochreiter & Schmidhuber, 1997).

In my thesis, I will use this type of RCNNs which has a ResNet as spatial feature extractor at the bottom and a 1d LSTM as spatiotemporal feature extractor at the top. This choice was made because the aim of the thesis was to assess and develop some basic and standard NNs for video analysis and the aim was neither choosing the more accurate NNs nor making the most biologically plausible model.

1.5.2.2 Recurrent Convolutional Neural Networks of Computational Neuroscientists

On the other hand, computational neuroscientists (Kubilius et al., 2018; Liao & Poggio, 2016) proposed RCNNs that are more biologically plausible and can fit the human behavioural performance such as the different accuracies and potentially even the different RTs in different experimental conditions. For instance, Liao and Poggio (2016) suggested that the three visual pathways can be modelled by multi-state fully recurrent neural networks. I will simply call these models fully recurrent neural networks (FRCNNs). Each visual stream can be represented by a cyclic graph G with vertices V and edges E: $G = \{V, E\}$. The vertices V are the set of layers. For instance, in the ventral stream $V = \{LGN, V1, V2, V4, IT\}$. RGC is not included because there no known connections from the visual cortex to the retina. However, RGC is the main bottom-up source to LGN which then influence the activity of all other areas. The edges E are a set of all connections between the layers V. The edges E include forward, backward (feedback) and lateral connections. Both the forward and feedback connections can further be simple or shortcut connections.

In FRCNN, the activation of every layer in a given timepoint t is a function of the activation of all layers, including itself, at the previous timepoint (t−1) by forward, backward, lateral connections with or without shortcuts.

1.6 Conclusions

Active action classification is a complex skill which require at least four visual subskills: object classification, action classification, visual perspective taking and on-line egocentric localization to constantly direct the current viewpoint to the target viewpoint on the moment-to-moment basis. These visual subskills have been widely studies in both humans and robots. However, they have been studies individually and there is a lack of knowledge about these visual subskills can interact for efficient active action classification. Towards filling that gab, the main aim of this thesis is to unveil whether and how both human and robotic observers are efficient active actions classifiers.

There are at least two types of experimental results that would confirm whether a specific type of observers are efficient active action classifiers. One, their action classification must be more accurate or faster in the active action classification conditions then in their corresponding baseline conditions. In the active action conditions, participants do proper active action classifications and therefore they can select their preferred viewpoints, by moving their viewpoints from the current positions to the preferred ones. In the baseline conditions, the observers cannot control the moves of their own viewpoints while they classify the actions of some actors. Some instances of baseline conditions are passive, in which viewpoint do not move at all, and random, in which the viewpoint change randomly. Two, the observers must select more often the efficient rather than the inefficient viewpoints in the active condition. To verify whether this is the case, we firstly need to identify the efficient and inefficient viewpoints for the studied types of observers in the passive condition. Secondly, we need to make statistical inferences about whether they do select the efficient viewpoints more often in the active condition.

The objectives of this thesis were seven. The first one was the selection of an appropriate action dataset to study active action classification in both humans and robots. For this objective, I reviewed the current action datasets which are only usable for passive action recognition studies. I highlighted what they lack to be suitable for active action classification studies and how my multi-view dataset overcomes the issue. So, my first research question was: which action dataset should I use to study both human and robotic active action recognitions? I addressed this research question in chapter 2.

The second objective was to discover the efficient and inefficient views for the action recognition of humans. The efficient views for the action recognition of humans are the ones from which the action classification of humans is more accurate or shorter (faster) than the other views. On the other hand, the inefficient views for the action recognition of humans are the ones from which the action classification is less accurate or longer (slower). Thus, the second research question was: which viewpoints are efficient for action recognition of the human observers and which ones are inefficient? The study in chapter 3 slightly addresses the efficient and inefficient views for humans by looking at the different action recognition accuracies and the RTs of human participants given the different starting views. However, this study's results were inconclusive because the views of the human observers changed within each trial; therefore, the view was not rigorously controlled. The experiment in chapter 4 more rigorously highlights the efficient and inefficient views for humans because all human participants of this experiment classified the same actions from different views with no view movements (NMs), and so their views did not change within the trials.

The third objective was to discover the efficient and inefficient views for the action recognition of computational models. The efficient views for computer models are the ones from which the action classification of these models is more accurate than the other views. On the other hand, the inefficient views for the robotic observers are the ones from where the action classification of these robots is less accurate than the other views. There were no reaction times of the models in this thesis because simulating RTs is out of the scope of this thesis. Therefore, to determine the efficient and inefficient views for robots, I only compared their action classification accuracies (and not their RTs) from all views. Sometimes, I also looked at the model's action classification loss (classification error), which is supposed to be lower from the efficient views and higher from the inefficient views. Thus, the third research question was: which viewpoints are efficient and which ones are inefficient for the action recognition of the robotic observers? The studies in chapters 5 and 7 identified the efficient and inefficient views for the action recognition of machines by showing the different action recognition accuracies of some basic models from different viewpoints with NMs. These chapters also examined whether the pattern of efficient and inefficient viewpoints for action recognition is the same in humans and robots.

The fourth objective was to verify whether humans' action recognition is more accurate and faster with active self-controlled view movements (SCMs) than with passive random view movements (RMs). The action recognition with SCMs is active since the view movements are selected by the human observers from a pool of possible view movements. On the other hand, the action recognition with RMs is passive because the view movements are not selected by the observers. Instead, the RMs are randomly sampled from a pool of possible view movements. Then, the fourth research question was: the action recognition of humans is more accurate and faster with SCMs than with passive RMs? The studies in chapters 3 and 6 directly addressed this research question with different methods.

The fifth objective was to examine whether humans select the efficient views more frequently than the inefficient views when they actively recognise actions. Consequently, the fifth research question was: do humans select the efficient views more frequently than the inefficient views during their active action recognition? The pilot study of Chapter 3 computed the frequencies of the selected views by the human participants at the last timepoints of the active action recognition trials with SCMs just before their action classifications. Chapter 3 also calculated the action recognition accuracies and RTs of humans in different starting views. Then, the study investigated the correlation of these view frequencies of the SCM humans with the accuracies and RTs of humans in the different starting views. If the SCM humans selected the efficient views more frequently, then the view frequencies of the SCM humans were expected to be positively correlated to the accuracies of humans in the different starting views and negatively correlated with the RTs of the humans in the different starting views. Since the viewpoints of this study moved within trials, the views were not rigorously controlled. Thus, the results of this study were not conclusive.

Chapters 4 and 6 solved this limit. Chapter 4 computed the view accuracies and the RTs of the NM humans whose viewpoints were locked and did not change within trials. Chapter 6 calculated the frequencies of the selected views by the SCM humans and correlated them with the view accuracies and view RTs of the NM human participants, previously found in chapter 4. If the SCM human observers of chapter 6 selected the efficient views more often, then the view frequencies of the SCM humans were predicted to be positively correlated with view accuracies and negatively with view RTs of the NM humans of chapter 4.

The sixth objective was to identify whether the action recognition of the active robots is more accurate than the passive robots. Thus, the sixth research question was: can action recognition of the active machines be more accurate than the passive machines? The study of chapter 7 answered this question. In the study, there were two types of active robotic observers and two types of passive robots. The two types of passive models were the NM and the RM models. The NM models were trained, validated and tested with only NMs, while the RM models were trained, validated and tested with only RMs. The SCM and RSCM models were the other two types of active models. The random and self-controlled view movement (RSCM) models were trained and validated with only RMs, but they were tested with only SCMs. The SCM models were trained for active action recognition with both RMs and SCMs, while they were validated and tested with only active SCMs. Chapter 7 investigated whether the SCM and RSCM models were more accurate in action recognition than the NM and RM models.

The seventh objective was to examine whether robots select the efficient views more often than the inefficient views when they actively recognise actions. The seventh research question was: do robotic observers choose the efficient viewpoints more frequently than the inefficient viewpoints when they do active action recognition? The study of Chapter 7 computed the frequencies of selected views by the SCM and RSCM robotic observers at the most accurate timepoints of their active action recognition trials. Chapter 7 also calculated the action recognition accuracies and RTs of NM models whose viewpoints were locked and could not change within trials. Finally, chapter 7 computed and correlated the frequencies of the selected views by the SCM models with the view accuracies and view RTs of the NM models. If the SCM computer models of chapter 7 chose the efficient viewpoints more often, then the view frequencies of these SCM models were expected to be positively correlated with view accuracies and negatively with view RTs of the NM computer models.

Efficient active action classification models have the potential of improving the quality of security surveillance, health monitoring and intuitive human computer interfaces.

2 A Novel Multi-View Action Dataset for Active Action Recognition

2.1 Introduction

The focus of this chapter is neither action recognition nor action detection methods. Instead, it is only about the most popular and modern action datasets. In this chapter, I am going to review the most popular and modern action datasets. Generally speaking, these are not appropriate for neither active action recognition nor active action detection. because at least one of the following conditions is true: they do not have multi-view videos (MVVs); they have a very few views up to three; the images are not made on-the-fly by a 3D simulator from the selected viewpoint in the 3D space. MVVs are videos showing the same event from different positions. In the case of datasets with no MVVs, the observers cannot select their viewpoints. Instead, the only viewpoint of a specific action is forced to the observer. In the case of MVV datasets with only a few views, there are not as many views as in the real 3D world for the observers to choose. To my best knowledge, there is no on-the-fly 3D simulator for action vision processing. To tackle this issue, I am going to introduce my own dataset which I named multi-view videos of human actions (MVVHA). MVVHA is an excellent dataset to study active action recognition.

For both action recognition and action detection studies, the ideal dataset characteristics are the following: Time distributed; large number of samples; large number of human activity classes; inclusion of person-only classes, person-object classes and person-person classes; balanced numbers of samples across classes; data variability; multi-labels per actor; individual action labels for all actors; multi-modality; multi-view. However, specifically for action detection studies, it is needed a dataset with 2 additional characteristics. These are spatial and temporal annotations of the action labels.

Let us start by describing the ideal dataset characteristics for both action recognition and detection studies. Each sample of the datasets should be time distributed for action recognition and action detection. Given that every action happens in a time interval, time distributed data is crucial to discriminate different actions, especially for the inverse actions like grasping an apple from the table and putting an apple on the table. In most of the datasets, the samples are RGB videos and thus they are time distributed images. However, the image datasets (Chao et al., 2018; Chao et al., 2015; Le et al., 2014; Ma et al., 2017) for action recognition or action detection have images as samples and thus they are not time distributed.

Human observers can easily recognise multiple actions that are performed by an actor at the same time or in different times. Thus, for advanced action recognition, an ideal goal is to recognise multiple actions of an actor at the same time or in subsequent moments. This is quite realistic because people can do more than one action at the same time such as can sitting, listening to someone, watching TV and eating or in different moments such as “climbing and then fall”, “pour water and then drink”. To do so, it is necessary a dataset that has multi-labels per actor, i.e. multiple action labels corresponding to the same actor. sSuch datasets can stimulate the implementation of advanced models that can also recognise multiple actions of an actor. AVA (Gu et al., 2018) and Multi-Moments in Time (Monfort, Ramakrishnan, et al., 2019) are datasets with this characteristic.

In real life, human observers may see multiple actors at the same time. Yet, they may understand what each actor is doing. Thus, another advanced goal in action recognition is to develop methods that can individually recognise different activities of multiple actors in a sample. For this reason, action recognition researchers need action datasets with individual action labels for all actors. The action datasets should have data samples of multiple actors doing different activities and each of these samples should have different action labels for different actors. Most action datasets have videos with only one action label either referring to only one actor or a group of actors. However, some modern datasets like AVA (Gu et al., 2018) and Multi-Moments in Time (Monfort, Ramakrishnan, et al., 2019) have individual labels for all actors.

It has well known in machine learning that models, such as NNs, perform better in a specific task if they are trained with a larger number of samples. For instance, some studies (Du et al., 2018, 2019) have theoretically and experimentally proved it. Both their mathematical theorems and their experiments highlight that the more the number of training samples, the less the loss of simple deep learning models is. They proved it for fully-connected neural networks, CNNs and RNNs. Thus, since most (if not all) modern action recognition and action detection methods rely on deep learning, action datasets with more samples are more appreciated by the researchers of the field.

Good action datasets should have many classes of human activities because the main goal of action recognition is not to develop methods that only recognize a few human activities. Instead, the aim is to build methods that can recognize many human activities and ideally all possible human activities. Therefore, action datasets should contain as many classes of human activities as possible. In some modern datasets such as Multi-Moments in Time, HACS and NTU RGB+D 120 (Liu et al., 2019), there are hundreds of human activities.

To cover most of the possible human actions, a dataset for action recognition or action detection should have 3 main types of action classes: person-only; person-object; person-person. Person-only classes involve “solo” activities. These are activities that, in general, do not influence anything in the environment of the actor. Some instances of person-only action classes are running, walking, jumping, standing and solo dancing. Person-object (or object manipulation) classes include actions of actors that interact with or manipulate an object. In this type of classes, the object may be necessary to recognise the activity. Examples of person-object actions are reading a book, riding a bike, writing a paper with a pen and so on. The person-person classes are the social actions that involve interaction with either another person or other people. For example, the person-person action classes can be shaking hands with someone, listening to someone, watching someone, talking to someone or hugging someone.

The numbers of samples in different classes must be approximately equal. Otherwise, the models may get high classification accuracy by just classifying all samples with the action class with more frequent labels and not by actually recognising the actions in the samples. So, to make sure that the models learn to recognise activities, the classes must be relatively balanced in the number of samples. For instance, the label frequencies are very unbalanced in AVA (Gu et al., 2018).

Ideally, a good dataset should have data variability. The action recognition and action detection models should learn to deal with different illuminations, different backgrounds, different actors, different movements, different poses, different actor orientations with respect to the viewpoint, different actor sizes, different actor distances, different clothes, and partial occlusion. Data variability, together with a large number of samples would avoid model overfitting and help the computer models to generalize the patterns of data.

Another ideal characteristic is the multi-modality data. Most action datasets are collections of no-audio videos or images. But some datasets like Moments in Time (Monfort, Andonian, et al., 2019), Multi-Moments in Time and AVA, contain videos with audio. Some datasets like Multimodal and Multiview and Interactive (M^2I) dataset (A.-A. Liu et al., 2016), NTU RGB+D (Shahroudy et al., 2016) and NTU RGB+D 120 (Liu et al., 2019) have videos and even 3D skeletons, depth maps and infrared sequences of human activities. Multi-modality models may gain action recognition correctness.

Finally, multi-view datasets are useful for several reasons. With multi-view samples, the models can learn to recognize or detect action from any position and, most importantly, we can study active action recognition and active action detection by designing experiments where the models select the next best views from where to look at the actions. The aim of the active action recognition studies would not only be action recognition, but it is also efficient selection of the next best viewpoints to improve the accuracy of action classification. In active action detection, there is a additional goal with respect to active action recognition. This is to localize in time and in space where the recognized actions are in the videos. Unfortunately, there are only a few multi-view datasets and these only have 2 (A.-A. Liu et al., 2016) or 3 (Liu et al., 2019; Shahroudy et al., 2016) views. In most datasets (Gu et al., 2018; Monfort, Andonian, et al., 2019; Monfort, Ramakrishnan, et al., 2019; Xu et al., 2017; Zhao et al., 2019), they sourced the videos from websites like YouTube. However, we should avoid sourcing videos from random websites because they do not have videos of the same activity from different viewpoints. MVVHA has 40 views of each human activity. To my best knowledge, MVVHA is so far the most appropriate dataset for active action recognition.

Let us now elaborate on the two ideal dataset characteristics that are relevant only for action detection. To learn how to spatio-temporally localize the recognised actions in videos, deep learning models need to be trained with a video dataset that has the spatial and temporal annotations of each action label. In the case of image datasets (Chao et al., 2018), there can only be spatial annotations. In case of video datasets, some have both spatial and temporal annotations like in AVA (Gu et al., 2018) and Actor-Action Dataset

(Xu et al., 2017), while other datasets such as HACS segments (Zhao et al., 2019) provide only the temporal annotations and not the spatial ones. In most datasets, the spatial annotations were done as boxes in the images or video frames. However, in the Actor-Action Dataset (Xu et al., 2017), they spatially annotated the actions in some images of their videos at the pixel level.

2.2 Action Datasets

Here, I will review the action datasets. I will evaluate them based on whether they have the important dataset characteristics described in the previous section. The action datasets can be in single-view and multi-view. The single-view datasets have data showing each action from only one view even if different actions are showed from different views. Multi-view datasets contain data that show every action from different positions.

2.2.1 Single-View Datasets

2.2.1.1 Moments in Time and Multi-Moments in Time

Moments in Time (Monfort, Andonian, et al., 2019) is a large video dataset of events (or actions). This dataset was designed for automatic event understanding. It contains over 1 million videos which are 3 seconds long. These videos include both the sequences of RPG images (Figure 2.1) and the audio allowing multi-modality action recognition. They even included videos where the actions such as “clapping” cannot be seen in the sequence of the video images but can only be heard through the video audio. Each video has a single event label. In the datasets, there are 339 action classes. There are a minimum of 1,000 videos for each action label or (class) in the dataset. In this way, the classes are relatively balanced. The actors that perform the action in the video can be humans, animals and

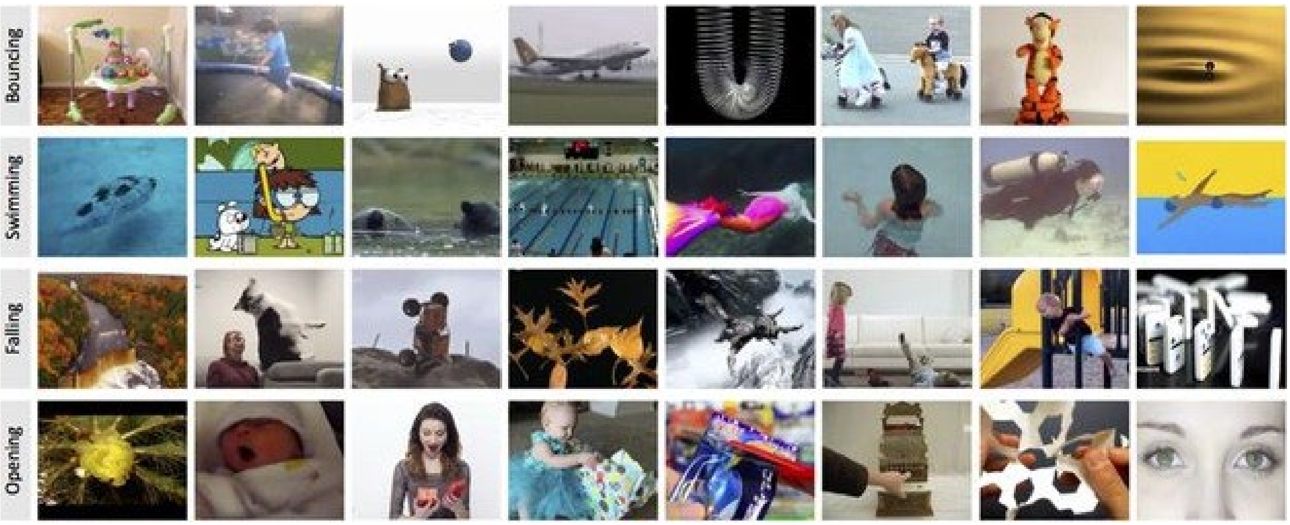


Figure 2.1. Same video frames in Moments in Time (Monfort, Andonian, et al., 2019).

objects.

There is a lot of intraclass variability that captures a dynamic event at different level of abstraction. For instance, opening doors, opening curtains, opening eyes, opening mouths, and even opening petals of a flower fall in the same action class “opening”. Additionally, there are numerous scenes and objects in the videos. In fact, the authors ran a 50-layer ResNet (He, Zhang, Ren, Sun, et al., 2016) trained on Places (Zhou et al., 2018) and a 50-layer ResNet trained on ImageNet (Deng et al., 2009) over 3 frames of each video. Then, they select the top-1 recognised scene and object for every video. In this way, they showed that the videos have 100% of the scene classes in the dataset Places and 99.9% of the objects in the dataset ImageNet.

Multi-Moments in Time (Monfort, Ramakrishnan, et al., 2019) is a newer version of Moments in Time. The main upgrade is that while each video in Moments in Time has only one action label, each video in Multi-Moments in Time can have one or more action labels. In other words, while Moments in Time has single action labels, the Multi-Moments in Time has videos with multiple action labels. In Multi-Moments in Time, they increased the number of labels to 2.01 million from 1.02 million videos. They slightly decreased the total of the action classes to 313, by removing 31 vague classes such as “working”, merging 37 classes into 20, and adding 22 new classes.

2.2.1.1.1 Strengths

Both Moments in Time and Multi-Moments in Time are a collection of videos (image sequences) and thus the data is time-distributed. This enables the observers to see the action developments through time. They contain a huge amount of samples with over 1 million videos. They also contain numerous action classes with a total of 339 label classes in Moments in Times and total of 313 classes in Multi-Moments in Time. There are person-only like walking, jumping, dancing and person-object actions such as “carrying”, “typing”, “repairing”, “painting” and more. It is unclear from the papers if they included person-person interactions such as hugging or hand shaking. Multi-Moments in Time has multiple labels per video, so there are different labels for different actors and different labels for the same actor within the same video. Moments in Time lacks this strength. The labels per class are relatively balanced as the minimum label count per class is over 1,000 while the average label count per class is around 2,200. Like other datasets sourced from websites like YouTube, Moments in Time and Multi-Moments in Time have a lot of data variability. The essence of their data variability is the presence of many different scenes and objects in different videos. Finally, the datasets are multi-modality as they have the image sequences and the audio of the videos.

2.2.1.1.2 Weaknesses

However, Moments in Time and Multi-Moments in Time have some limits. First, Moments in Time has only one label for each video. Therefore, there are not individual action labels for all actors in multi-actor videos and each actor does not have multi-labels within the same video. This is not a weakness in Multi-Moments in Time which has multi-labelled

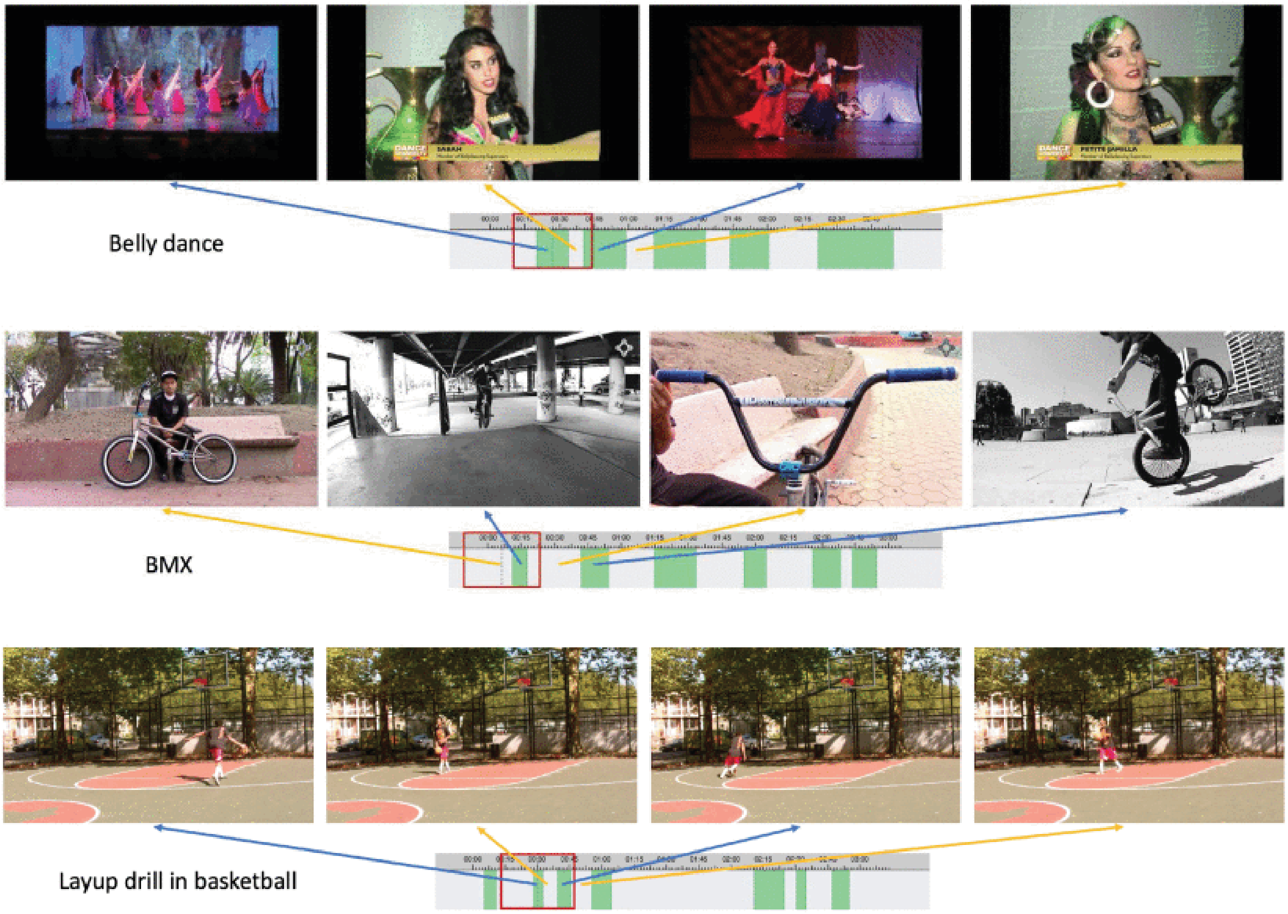


Figure 2.2. Examples of segment annotations in HACS (Zhao et al., 2019).

videos. It is not multi-view given that there not multi-videos from different viewpoint showing the same event. The lack of the latter characteristic makes the dataset not suitable for active action recognition where the observers can select the viewpoint where they can watch the actor from. Lastly, the dataset does not contain both spatial and temporal annotations of the action labels.

2.2.1.2 HACS

Human Action Clips and Segments (HACS) is a recent video dataset (Zhao et al., 2019). The videos were sourced from YouTube. It has two types of datasets: HACS Clips and HACS segments (Figure 2.2). HACS Clips are short videos and each of them has only an action label. HACS Segments have both the action labels and the action temporal boundaries. HACS clips consist of 1.5M short videos of 2 seconds that were sampled from the 504K untrimmed YouTube videos, while HACS segments includes 139K action segments in 50K untrimmed videos. The untrimmed videos of HACS segments are long up to 4 minutes. Both types of action analysis benchmarks contain the same 200 classes of actions. In such videos, there is a lot of realistic data variability in illumination, viewpoints, video quality, background and actor ethnicity. Figure 2.2 highlights some HACS segments in 3 different videos. To speed up the annotation process, at first, they used some automatic methods to label the videos and annotate the action temporal boundaries. However, these labels and annotations were confirmed by human users through a user-based software. They provided to the human annotators a guideline which contains the action definitions to reduce ambiguity of actions.

2.2.1.2.1 Strengths

HACS haves several strengths. Their samples are videos, so they are time distributed data. This dataset is one of those that have the largest total numbers of samples. There are 1.5M videos in HACS clips. There are many action classes: 200 classes. They have person-only, person-object and person-person classes. They have a lot of data variability.

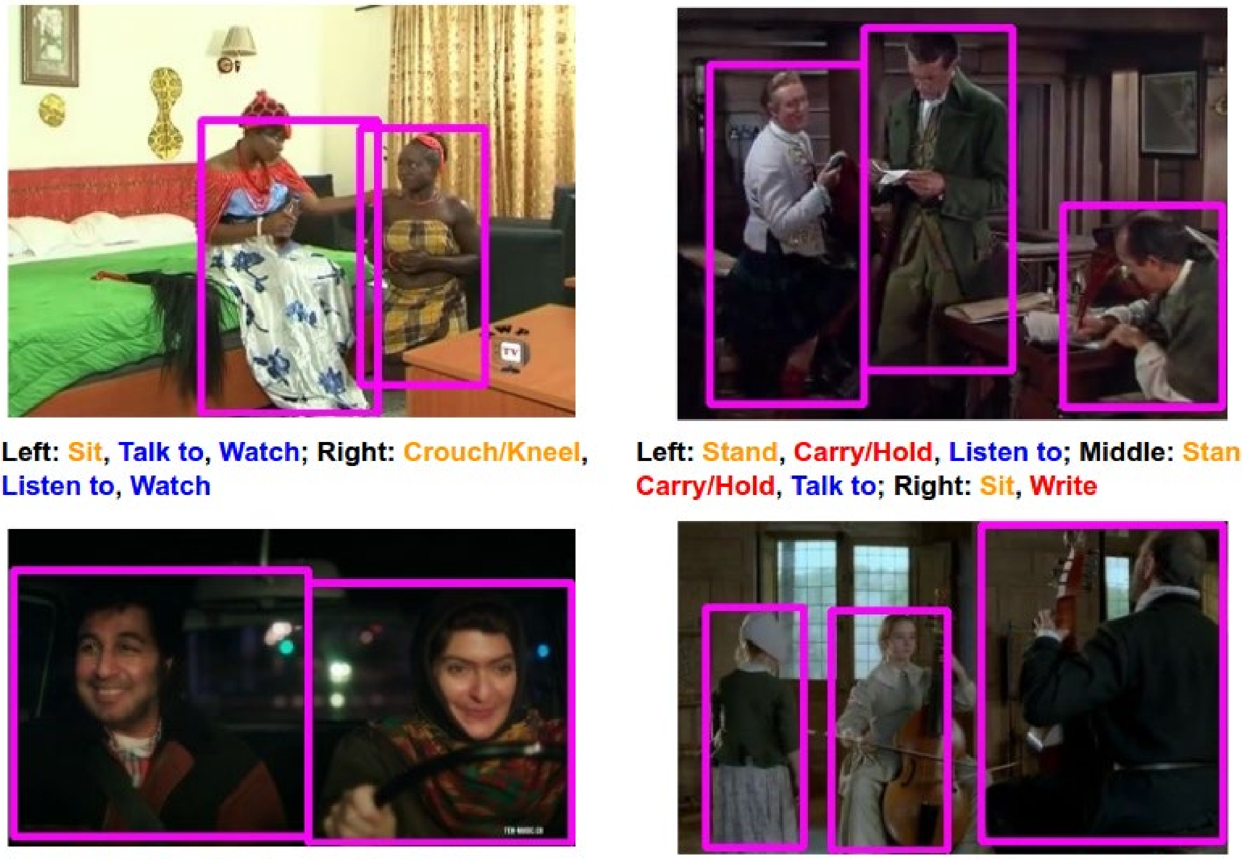
2.2.1.2.2 Weaknesses

However, it has some weaknesses. HACS Clips is suitable only for action recognition and not for action detection because it does not have spatiotemporal annotations of the actions which make the dataset unsuitable for action detection. However, HACS segments do not have the spatial annotations, but have the temporal annotations. Therefore, researchers can use HACS Segments for temporal-only action detection. In the paper, they do not mention whether the number of samples are approximately balanced across classes. The samples only have one label. Thus, they do not have multiple labels for each action performed by an actor at the same time or at different times. In addition, each video does not have individual labels for each actor in the video, but it only has a collective label for all actors or a label for only one actor. They are not multi-modality data as it only contains videos. They do not have MVVs. In fact, they do not have different videos of the same activity from different positions. This makes them not suitable for an active action recognition as the action recognition models cannot select the view from where to look at the action.

2.2.1.3 AVA

Spatio-Temporally localized Atomic Visual Actions (AVA) is a dataset of videos with spatial and temporal annotations of each action (Gu et al., 2018). AVA consists of 430 videos of 15 minutes with 1.58 million action labels that are localised in space and in time. The 430 videos were sourced from the 15th to the 30th minute time intervals of 430 movies. It contains 80 classes of actions. There are multiple action labels corresponding to one person in the same time interval. To make the task even more challenging, in the videos there are often multiple actors at the same time and each of these actors have their own multiple action labels with spatial and temporal annotations. In fact, every person in the videos is localized with bounding boxes in each frame per second. Then, multiple action labels are attached to each bounding box (or actor). Finally, each action label has the temporal start and end points.

There are three main types of action labels: 14 pose labels regarding to the actor pose such as sitting, walking, standing; 49 person-object interaction labels corresponding to interaction with objects like carry, write; and 17 person-person interaction labels that concern to interaction with other people like talk to someone, listen to someone and watch. Every person in a frame is always labelled with a pose label and they may have additional



Left: Sit, Talk to, Watch; Right: Crouch/Kneel, Listen to, Watch

Left: Stand, Carry/Hold, Listen to; Middle: Stand, Carry/Hold, Talk to; Right: Sit, Write

Left: Sit, Ride, Talk to; Right: Sit, Drive, Listen to

Left: Stand, Watch; Middle: Stand, Play instrument; Right: Sit, Play instrument

Figure 2.3. The bounding box and action annotations in sample frames of the AVA dataset (Gu et al., 2018). Each bounding box is associated with 1 pose action (in orange), 0-3 interactions with objects (in red), and 0-3 interactions with other people (in blue).

object interaction labels up to 3 and additional person-person interaction labels up to 3 (Figure 2.3). Thus, each person always has at least one action label up to 7.

They labelled the frames at 1 Hz and made 900 keyframes per movie (15 minutes x 60 keyframes per minute = 900 keyframes) and 387,000 keyframes in the whole dataset (900 keyframes per movie x 430 movies = 387,000 keyframes). Each person is linked to the consecutive keyframe to provide the temporal sequences of the action labels. Figure 2.3 shows some keyframes of the dataset.

2.2.1.3.1 Strengths

The samples of AVA dataset are videos, so they are time distributed images. It has both spatial and temporal annotations of all actions, then it is appropriate for action detection. It has individual labels for all actors in the videos and additionally each actor in the videos have multiple actions labels at the same time or in different moments. It has 1.58 million labels in 6450 minutes of video (430 videos times 15 minutes per video) so it has a massive number of samples. It has a lot of action classes (80) and these include person-only (pose), person-object and person-person classes. They searched the movies by top actors of different nationalities. Therefore, there is a lot of data variability in illumination, culture, clothes, actors, video quality, background and so on.

2.2.1.3.2 Weaknesses

The classes do not have balanced number of labels. They said that this is how it should be because the action distribution of a dataset should have a realistic distribution which is unbalanced. In other words, some categories of actions like walking happen more often in real life than others such as punching. The action recognition methods should learn the realistic distribution of the actions. However, the dataset only has a very few labels (about 20 to 40) for some action classes which are not enough to train, validate and test deep leaning models. It is slightly multi-modality as the videos contain the audios. However, they do not contain any other types of data like depth maps or 3D skeletons of the actors. Another limit of the dataset is that the videos are single-view. In fact, they show each action from one single view. Given that AVA is not multi-view it is neither appropriate for active action recognition nor active action detection.

2.2.2 Multi-View Datasets

2.2.2.1 Multimodal and Multiview and Interactive (M^2I) dataset

M^2I dataset is one of the first multi-view datasets with only two views. It also a multi-

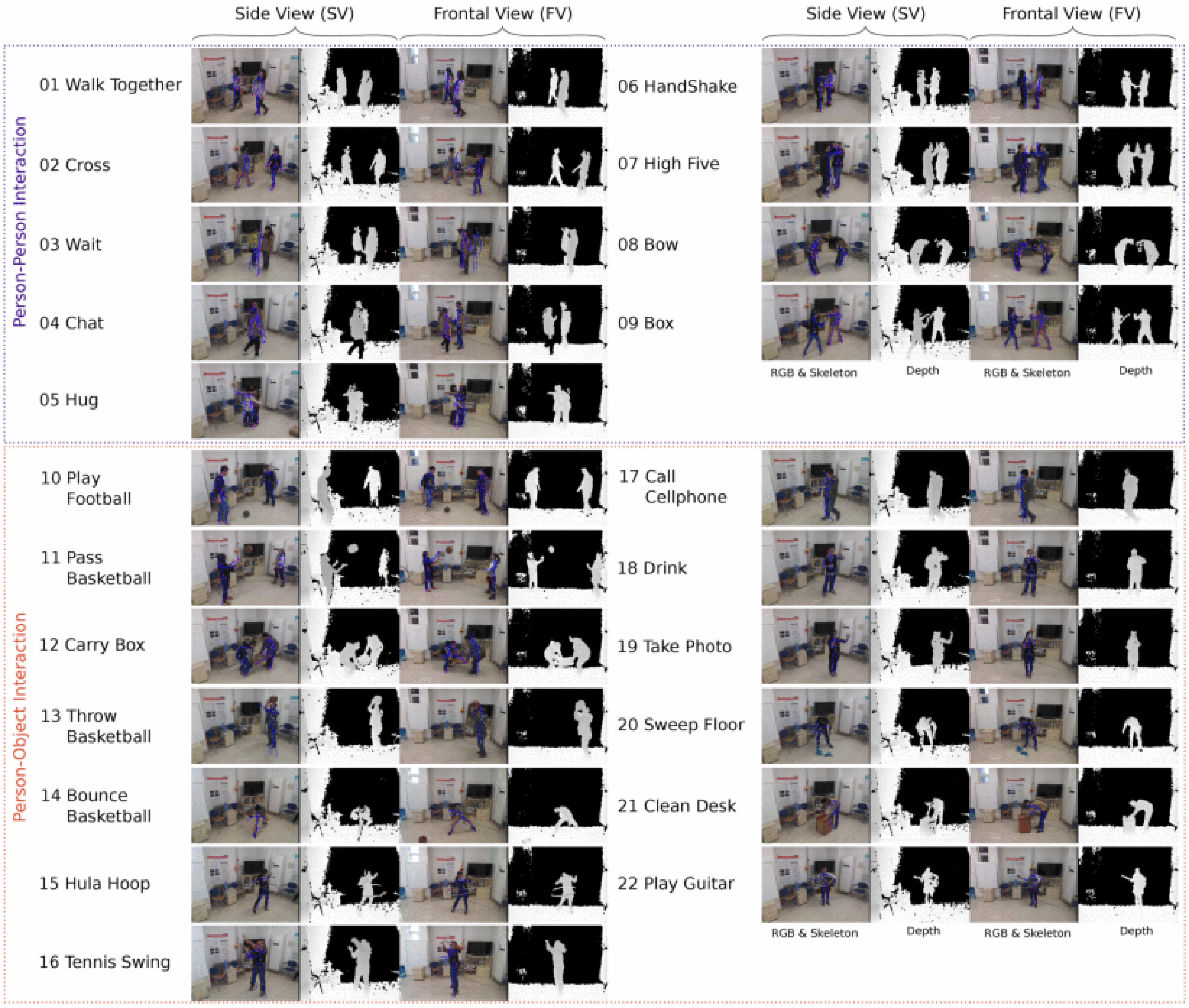


Figure 2.4. Image samples and action categories of the M^{2}I dataset (A.-A. Liu et al., 2016).

modality dataset of human activities as each sample has three types of data: RGB videos, Depth maps; skeletons. The shapes of both the RGB videos and the depth maps are 320 x 240. The skeletons are 3D coordinates of 20 body joints data. All three types of data are recorded at 30Hz. There are 22 classes of actions. The dataset includes atomic actions, person-person interactive actions and person-object interactive actions (Figure 2.4). Each action is performed twice by 20 person-person or person-object pairs and each time the action was simultaneously recorded from two different views which were frontal view and side view. Therefore, there was a total of 1,760 videos in the dataset (22 actions x 20 pairs x 2 views x 2 runs). All videos were recorded in the same room with the same background even if they slightly varied the illumination.

2.2.2.1.1 Strengths

The dataset is time distributed data at 30Hz. The action categories include person-only, persons-person, and person-object actions. The number of samples are perfectly balanced across classes of actions. It is multi-modality as it has 3 types of data: RGB videos, depth maps and skeletons.

2.2.2.1.2 Weaknesses

The dataset does not have a lot of samples (1,760). It only has 22 classes of actions which are quite enough, but they are not very many. The variability of data is great even if they slightly varied the illumination and there used 22 different actors, the background has always the same objects and same wall. The samples only have one single-label for all actors and, therefore, there are neither multi-labels for a single actor nor individual labels for all actors. The dataset is multi-view, but it is relatively good for active action recognition because it only has two views which are not representative of all potential viewpoints around the action. Finally, it does not have any spatio-temporal annotations, so it is not adequate for action detection.

2.2.2.2 NTU RGB+D and NTU RGB+D 120

NTU RGB+D is a multi-modality and multi-view dataset (Shahroudy et al., 2016). Overall, the dataset contains 56,880 RGB+D samples. Each sample has 4 data modalities: RGB videos, depth maps, infrared (IR) frames and 3D skeletons of the people performing a specific action. The depth maps and the IR sequences have a resolution of 512 x 424 pixels, while the resolution of the RGB videos is 1920 x 1080. The 3D skeletons consist of

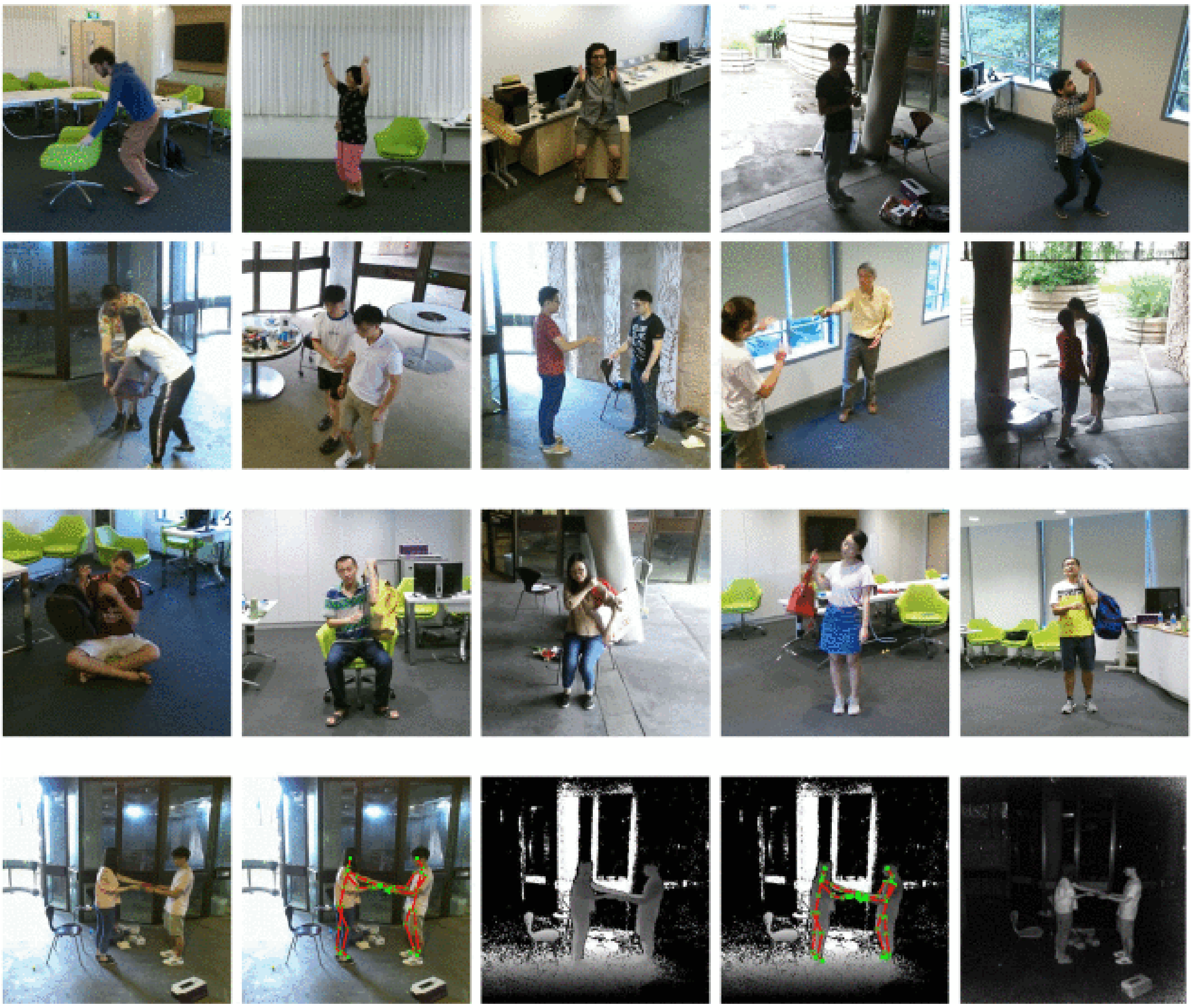


Figure 2.5. Sample frames of the NTU RGB+D 120 dataset (Liu et al., 2019). The first two rows show the variety in human subjects, camera views, and environmental conditions. The third row depicts the intra-class variation of the performances. The last row illustrates the RGB, RGB+joints, depth, depth+joints, and IR modalities of a sample frame.

3D points of 25 major body joints of the actors/subjects for each video frame. There are 60 classes of actions. They hired 40 subjects/actors for the dataset. There is a lot of variability in the subjects, in gender, height and age. Their ages ranged from 10 to 35 years. By looking at the camera views, the videos were taken from 80 points of views. They recorded 3 videos of the same action from 3 different locations. In each setup, the three cameras were placed at the same height but at three different horizontal angles: 0°, ±45° and ±90°. The participants were asked to do the same action twice: once towards the left camera and once towards to the right one. Therefore, for each setup, they made 6 videos from 5 different views: 2 front views, one 90° left side view, one 90° right side view, one 45° left side view and one 45° right side view. The three cameras were assigned three different labels based on their position angles: Camera 1 is the 45° view, Camera 2 is the front view (0°), Camera 3 is side one (90°). They made 17 different camera setups by varying the distance and the height of the camera views to produce more views. They made with the datasets 2 types of classification benchmark for action recognition by splitting the dataset in two different ways. One is cross-subject classification benchmark where the data is split based on the subjects, such that the test data come from the data of 20 subjects and the training data is the data of the other subjects. In other words, in the cross-subject classification, the NNs are trained on some subjects and tested on other subjects. The second classification benchmark is cross-view: the NNs are trained on Cameras 2 and 3 (front view, 90° left side view and 90° right side view) and tested on Camera 1 (45° left side view and 45° right view).

The dataset NTU RGB+D 120 (Liu et al., 2019) is an extension of the dataset NTU RGB+D (Shahroudy et al., 2016). In the new version, the number of samples is approximately as twice as the previous one: it has 114,480 RGB+D samples instead of 56,880. The human action categories are also extended from 60 to 120. The multi-view dataset has now 155 camera viewpoints (instead of 80). The actors are 106 (instead of 40) and have more data variability in gender, height age and culture. The height ranged from 1.3m to 1.9m, the age range is 10 to 57 (instead of 10 to 35) and they come from 15 different countries. To add further data variability to the dataset, they used 96 different backgrounds by varying the scene illumination. Finally, they nearly doubled the camera setups: now they are 32 (instead of 17). Therefore, they also increased the camera viewpoints. In each camera setup, there are still three cameras recording the same action from the same height but from different angles: 0°, ±45° and ±90° (Figure 2.5). For each setup, they again made the same 6 videos from 5 different views like in the previous version, by recording the action twice: one time the action is towards the left camera and one time it is towards the right one. Regarding the type of data, they collected the same 4 data modalities of the first version: RGB videos of 1920 x 1080; depth sequence and IR sequence of 512 x 424; 25 3D locations of the major body joints. They made two types of action classification evaluations: in the cross-subject evaluation they trained NNs with the data of some actors and tested them on the data of other unseen subjects; in the cross-setup evaluation, they trained NNs on some camera setups and tested them on other setups. Overall, the cross-setup classifications of those NNs are more accurate than the cross-subject one. Additionally, the Body Pose Evolution Map scored the best accuracy in both types of evaluations with a cross-subject accuracy of 0.646 and a cross-setup accuracy of 0.669.

2.2.2.2.1 Strengths

NTU is multi-modality as provide RGB videos, depth maps, IR sequences and 3D skeletons. All the four modalities of data are time distributed. They have a massive number of samples, and the action classes have balanced numbers of samples. NTU includes person-only, person-object, person-person and person-object-person action. Examples of person-object-person actions are “wield knife towards other person” and “hit other person with object”. They put some degree of efforts on adding some data variability such in actor culture, in actor age, illumination, views. However, apart from illumination, there is not a lot of data variability in background as all recordings were done in same lab. Finally, it is slightly multi-view because there 3 videos from 3 different positions displaying the same activity. This can be used for active action recognition as the NNs can choose from which one of the three positions to look at the activity. However, 3 positions are quite a few compared to the potential number of views which an action can be seen from.

2.2.2.2.2 Weaknesses

NTU only provides one label per sample, then it does not have individual labels for each actor in the samples. Furthermore, every actor is not labelled with multiple actions like “standing” and “shaking hands”. The dataset is only suitable for activity human recognition and not for human activity detection as it does not provide the spatio-temporal annotations of the action labels.

2.3 Multi-View Videos of Human Actions (MVVHA)

As stimuli of the action recognition task of both human and robotic participants, I used my own dataset of multi-view videos of human actions (MVVHA). I made several versions of the datasets with very little differences. I will describe here the details of the most updated version r.3.5 that was designed for computer vision experiments. The “r” in r.3.5 stands for robotic. Later, I will talk about the other versions and their differences with the dataset version r.3.5. The dataset version r.3.5 contains 24,570 MVVs and each MVV is a collection of 40 single-view (ordinary) videos (SVVs) showing the same actor performing the same action from 40 different viewpoints. Therefore, there are nearly a 1 million (982,800) SVVs in this dataset version. Figure 2.6 shows a multi-view frame of an MVV. All videos are action-labelled and that make it suitable for action classification. The classes of actions are perfectly balanced. In fact, there are exactly 140,400 SVVs (3,510 MVVs) in each class of actions.

These videos were made by rendering 3D animations by Blender (http://www.blender.org) from different viewpoints. I downloaded different 3D actor animations and different 3D actor bodies from Mixamo’s website (https://www.mixamo.com). Next, I uploaded one 3D

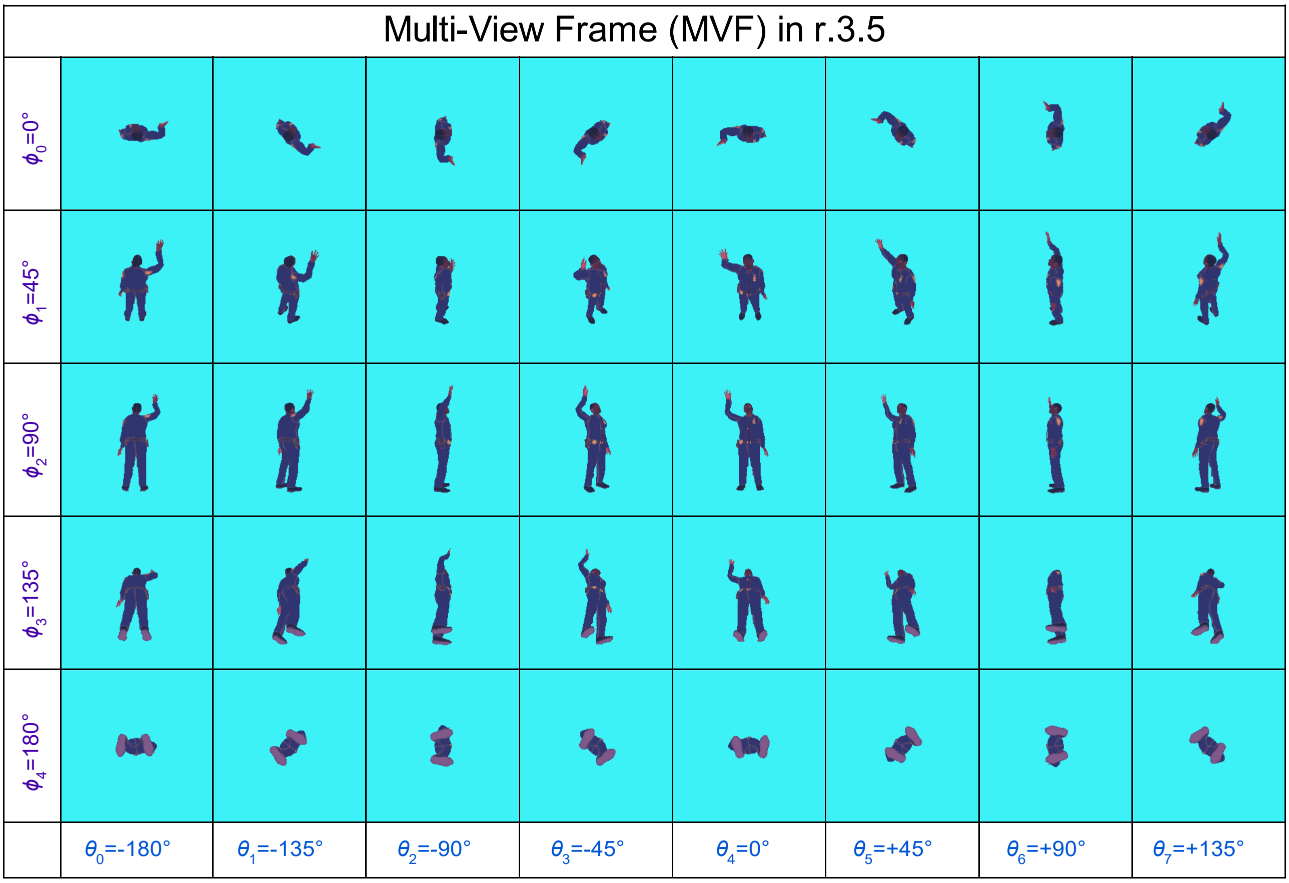


Figure 2.6. A multi-view frame of a multi-view video (MVV) in the version r.3.5. Each MVV is a collection of single-view (ordinary) videos (SVVs) that show the same actor executing the same action from different viewpoints. The viewpoint of every SVV is defined mathematically in 3D space by 3 parameters: ρ, θ, ϕ. ρ is the distance between the viewpoint and the actor's hips, θ is the azimuth angle of the viewpoint around the actor and the ϕ is the elevation angle of the viewpoint with respect to the actor's hips.

animation and one 3D actor at the time into Blender. Following that, I animated the 3D actor with the 3D animation and made a MVV by rendering the animated 3D actor from different viewpoints at different time points. I automated both Mixamo downloading and Blender rendering with Python code. I scripted the download with Selenium (https://selenium-python.readthedocs.io/index.html), a Python library which provides functions to drive internet browsers like Firefox and Chrome.

2.3.1 Spherical Coordinates

All viewpoints P of the MVVs are defined mathematically by a spherical coordinate system

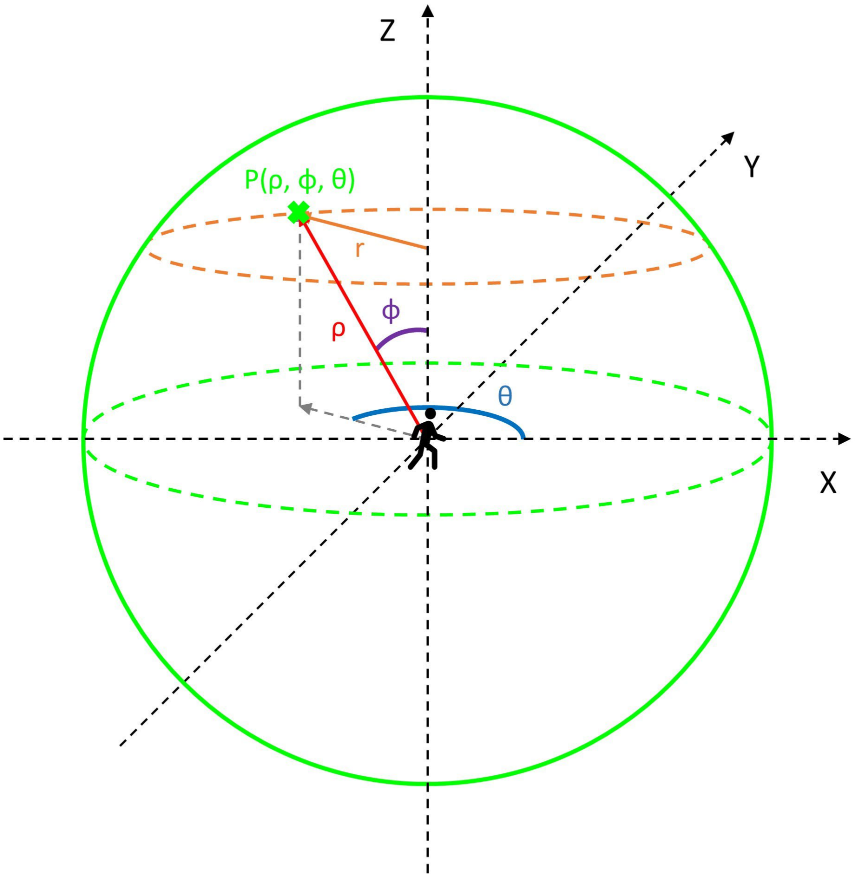


Figure 2.7. We defined the cordinates of the viewpoints through spherical coordinates. The reason for this choice is that all viewpoints of videos fall over the imaginary sphere (green cycle) that has its centre $P_0(x_0, y_0, z_0)$ in the middle of the hips of the actor. In the spherical cordinates, each point P on the imaginary sphere can be represented with three parametars: ρ, ϕ, θ, where ρ is the distance of any point on the imaginary sphere from the centre P_0, ϕ is the angle from the positive z-axis to vector ρ and θ is the angle from the positive x-axis to the projection of ρ on the plane XY. In addition, in the figure, r is the shortest distance between the z-axis and the viewpoint P. Note that the actor body faces towards the positive x-ax*is.*

with 3 parameters ρ, ϕ and θ (Figure 2.7). All views are equally distant from the actor's hips by the distance ρ. Therefore, all views fall on the surface of an imaginary sphere with centre $P_0(x_0, y_0, z_0)$ at the hips of the actor. The angle ϕ describes the elevation of the viewpoint with respect to the actor' hips, while θ is the angle around the actor' body. The 3 parameters had the following constraints:

$$\rho \geq 0$$

$$0° \leq \phi \leq +180°$$

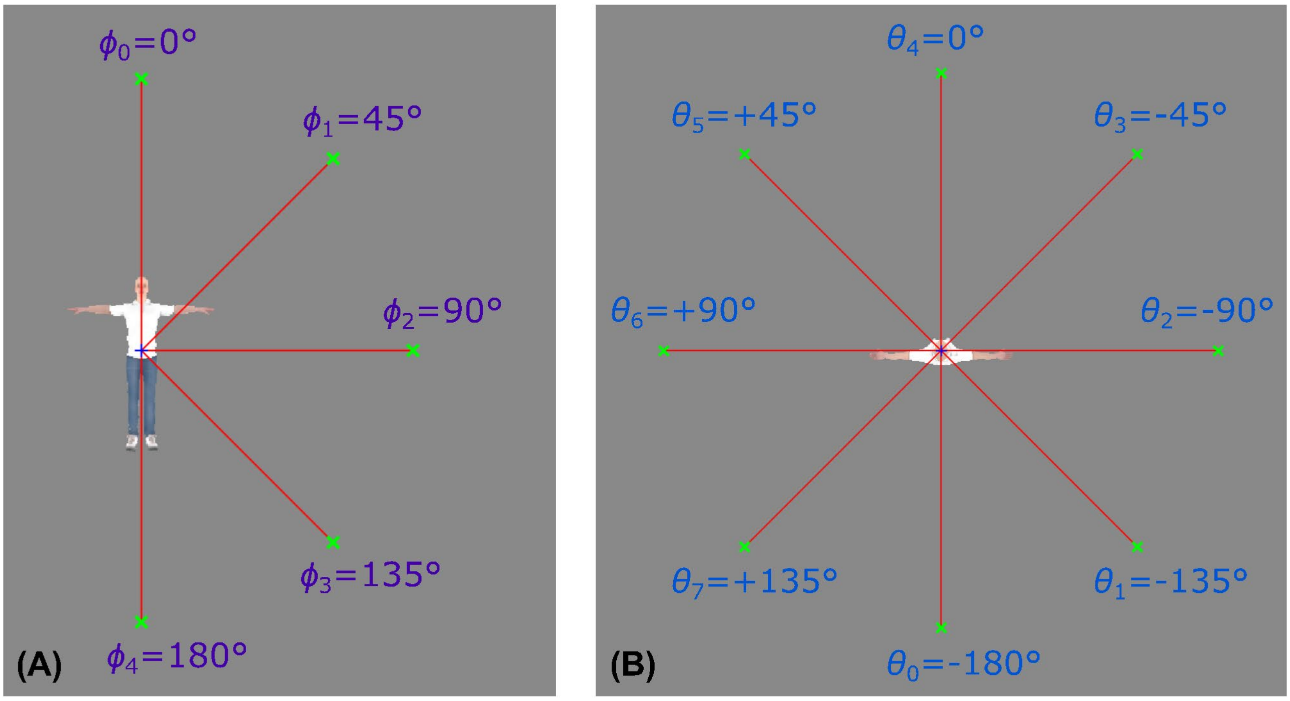


Figure 2.8. (A) shows the 5 angles ϕ which can either be 0°, 45°, 90°, 135° or 180°. A viewpoint with ϕ=0° and ϕ=45° is higher than the hips. For ϕ=90°, the viewpoints are as high as the hips with respect to the z-axis. The viewpoints with ϕ=135° and ϕ=180° are lower than the hips with respect to the z-axis. (B) shows all different angles θ of the viewpoints which can either be −180°, −135°, −90°, −45°, 0°, +45°, +90°, +135°. A viewpoint with θ=−180° shows the back of the actor, while a viewpoint with θ=0° is in front of the 3D person. Furthermore, a viewpoint is on the right side of the actor if −180°<θ<0°, whereas it is on the left if 0°<θ<180°.

$$-180° \leq \theta < +180°$$

As highlighted in Figure 2.8(A), for ϕ=90°, the views have the same height of the hips along the z-axis. For 0°≤ϕ< 90°, the viewpoints are higher than the hips, while, for 90°<ϕ≤180°, the viewpoints are lower than the hips. With respect to the angle θ, the actor faces towards the positive x-axis and, as shown in Figure 2.7, the views in front of the 3D person are defined by θ=0°. Furthermore, Figure 2.8(B) shows that, for positive θ or for 0°<θ<180°, the viewpoints are on the left of the actor, whereas the viewpoints with negative θ, −180°<θ<0°, are on the right side of the 3D actor. For θ=−180°, the views are exactly on the back of the actor. The corresponding cartesian coordinates of the points P is defined as:

$$P(x, y, z)$$

where

$$x = x_0 + r\cos\theta$$

$$y = y_0 + r\sin\theta$$

$$z = z_0 + \rho\cos\phi$$

and

$$r = \rho\sin\phi$$

2.3.2 Data Augmentation and Random MVV Parameters

I augmented the MVVs by mirroring each 3D animation in 3D space. In this way, I made 2 mirror conditions: mirror_00 and mirror_01. The 3D animations were not modified in mirror_00, while the 3D animations were mirrored in the 3D space in mirror_01, by inverting the left and the right sides of the original 3D animations, by inverting the left and the right sides of the 3D animations. For example, if the actors in a given animation point with the left hand in mirror_00, they do the same movement with the right hand in condition mirror_01. Mathematical speaking, if a 3D point P_{m0} of an animated actor in the condition mirror_00 was defined as:

$$P_{m0} = (x, y, z)$$

The coordinates of its corresponding 3D point P_{m1} of the same animated actor in mirror_01 was defined as

$$P_{m1} = (x, -y, z)$$

Figure 2.9 shows some corresponding frames of mirror_00 and mirror_01 in r.3.5. Note that they are not exactly symmetric because I randomized some scene parameters for each MVV.

I randomized some parameters of the scene such that there was some realistic data variability which prevents overfitting deep learning models during training. In particular, I randomly sampled from a uniform distribution the Mixamo parameters of each animation which slightly modify the poses and movements of the 3D animations. The type and number of Mixamo parameters were different for each 3D animation. I also added some gaussian noise to the dimensions of the actor body in all 3 cartesian dimensions x, y and z. I also uniformly randomized the colour of the background, the colour of the light, the light intensity and some extra colour on the surface of the actor body. These noisy parameters were kept constant for each video of the same MVV but varied between different MVVs.

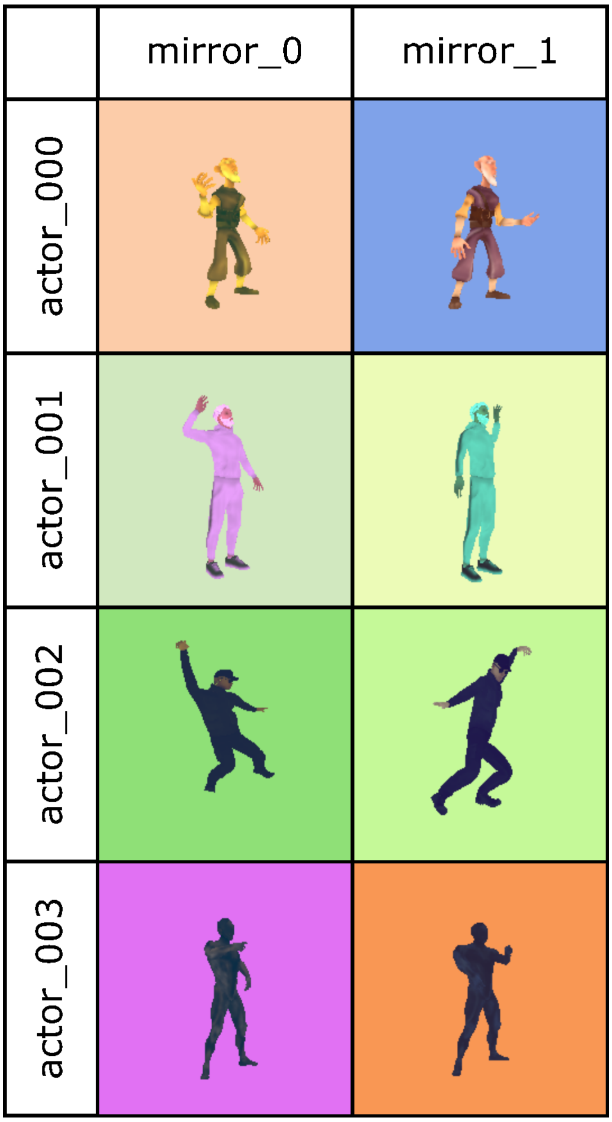


Figure 2.9. These are frames of some videos in the computer vision version r.3.5 of the dataset. They show 4 out 65 actors. These are the corresponding frames of the 2 mirror conditions for the same 3D actor animations, the actors, from the same viewpoint (ϕ_2=90°, θ_3=-45°), and at the same timepoint. The mirrored images are not exactly the same because some random parameters were defined for each MVV like the background colour, the light colour, light intensity, the scale of the actor body.

2.3.3 Dataset Dimensions

The whole dataset can be thought as a large array with 11 dimensions (or variables).

These dimensions can be split into between-SVVs and within-SVVs. The between-SVVs dimensions are class of actions, actor, 3D animation per class, mirror, distance ρ, angle θ, angle ϕ. On the other hand, the within-SVV dimensions are time, colour channel, x pixel, y pixel.

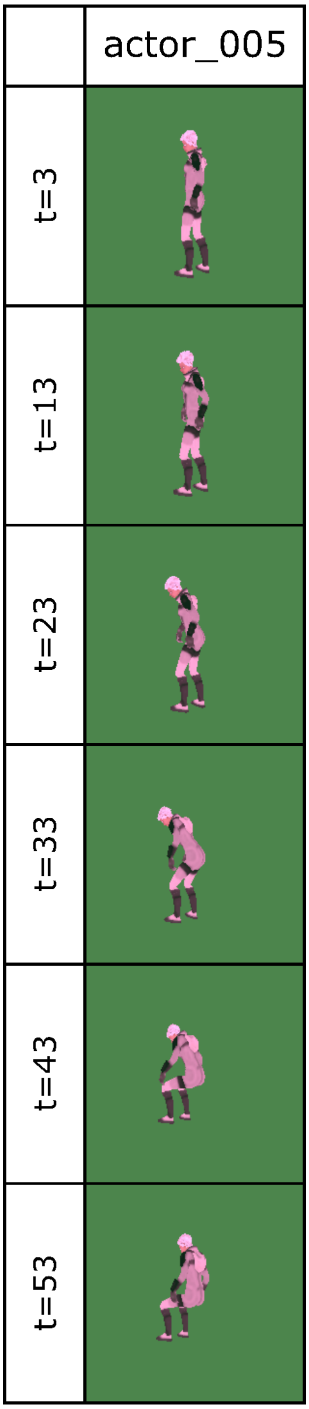


Figure 2.10. It shows 6 frames out of 60 of a SVV in the dataset version r.3.5. The class of the action that the actor in this SVV does is “Sitting Down” (Standing-To-Sitting).

Let’s look at the conditions of the between-SVVs dimensions. There are 7 different classes of actions: dancing, discussing, sitting down (standing-to-sitting), standing up (sitting-to-standing), falling down, pointing, waving. Each video has an action label which only belongs to one of the total 7 classes. There are 65 actors. In Figure 2.9, there are some images of the first 4 actors. There are 27 3D animations per class and a total of 189 3D animations for all classes. There are two conditions for the mirror dimension (Figure 2.9): mirror_00 and mirror_01. There is only 1 viewpoint distance ρ which was set to 7 in Blender. There are 8 azimuth angles θ. These are: −180° (back), −135° (right-back), −90° (right), −45° (right-front), 0° (front), +45° (left-front), +90° (left), +135° (left-back). The 8 angles θ are shown in Figure 2.6 and Figure 2.8(B). There are 5 viewpoint elevation angles ϕ: 0° (top), +45° (middle-top), +90° (middle), +135° (middle-bottom), +180° (bottom). They can be seen in Figure

2.6 and Figure 2.8(A).

Talking about the within-SVVs dimensions, there are 60 time points or frames for each SVV and since the original duration of every animation is 2 seconds, this dataset version has a refresh rate of 30Hz. Figure 2.10 provides a sight of time by showing the 6 frames out of 60 in an SVV. The timepoints of the 6 frames are 3, 13, 23, 33, 43, 53. The last three within-SVVs dataset dimensions are the single image dimensions. Each image's shape is 3 x 224 x 224, i.e. the frames are 224-pixel squares encoded with 3-channel RGB values.

Each between-SVV dimension is condition-labelled. Therefore, each video has a label of action class, a label of actor, a label of 3D animation within a class of actions, a label of mirror condition, a label of rho, a label of angle θ, a label of angle ϕ. In this way, the conditions of each between-SVV dimension can potentially be used as independent variable and as classes of some recognition tasks. For example, we could design a study with MVVHA to test whether action and actor recognitions are affected by the angle θ and the angle ϕ.

The seven action classes in MVVHA are similar. For instance, discussing, pointing and waving comprise similar body movements. Sitting down and Standing up are the same, except for the inverted time. Additionally, dancing and falling down may sometimes be similar. By assuming that action recognition and object recognition may be analogous, I selected these similar action classes because I expected that the similarities across actions would make the recognition performance viewpoint-dependent. The object recognition performance is viewpoint-dependent when the recognition accuracy or RT significantly vary in different viewpoints. On the other hand, the recognition performance is viewpoint-independent when the accuracy or RT are not affected by the change in viewpoint. Many researchers (Hayward & Williams, 2000; Leek & Johnston, 2006; Tarr &

Hayward, 2017; Tarr & Pinker, 1990; Tjan, 2001) have shown that the object recognition performance is viewpoint-dependent when the recognition task involves discriminating objects whose shapes are very similar to each other, while the object recognition performance is viewpoint-independent when distinguishing objects with very different shapes. The introduction of chapter 4 includes a broad literature review of the viewpoint-dependent and viewpoint-invariant object recognition performances. By choosing these similar action classes, I wanted to design a test where participants show viewpoint-dependent action recognition performance. In this way, I could examine whether they select the viewpoints efficiently during proactive action classification in that viewpoint-dependent test. It would not be possible to highlight the efficiency of the viewpoint selection of the participants in a test where participants show viewpoint-invariant action recognition performance because any viewpoint selection of the participants would not affect the performance anyway.

2.3.4 Strengths

MVVHA is time-distributed data. It includes many samples: there are 3,2 million SVVs. The action classes are balanced in terms of number of samples. Most importantly, it is multi-view. It contains 60 views which are far more than the views of any other action datasets. This makes it ideal for active action recognition.

2.3.5 Weaknesses

Our action dataset has some limits. It only has 7 classes of actions. Therefore, it does not contain a lot of activity classes. There are not person-object and person-person classes. In each video, there is only one actor doing the action and nothing else in the background. Thus, even if I added some data variability to the samples like 65 different actors, mirrored

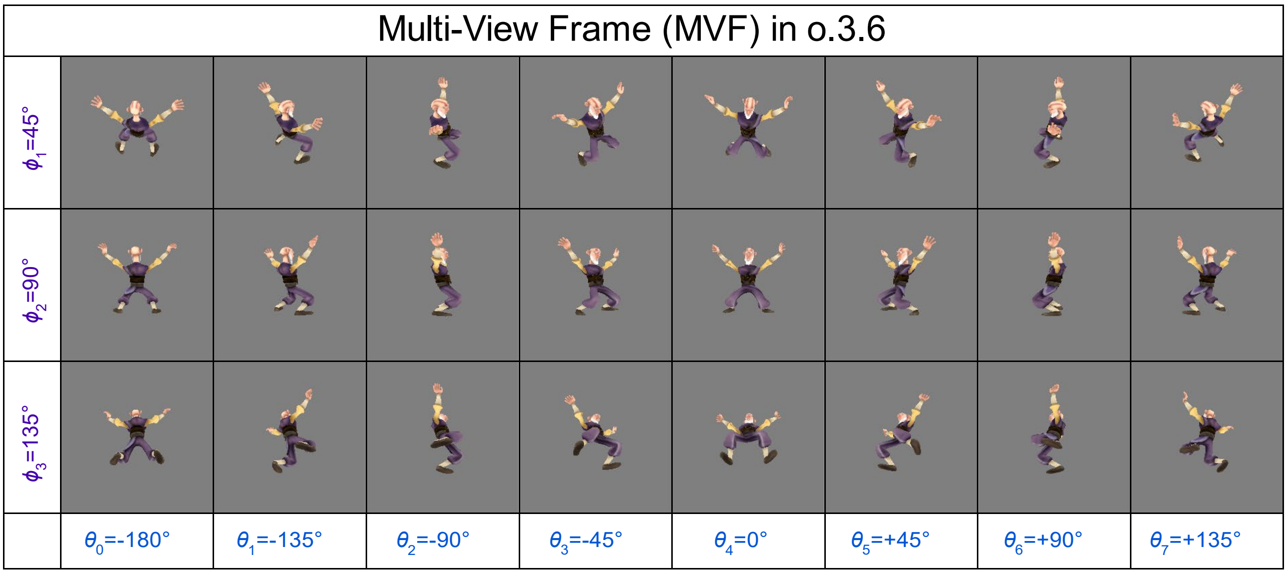


Figure 2.11. A multi-view frame (MVF) of a MVV in the online psychology version o.3.6. The class of action is Dancing in this MVV.

3D animations, different views, random background colour, the background is always empty. No sample has multiple labels. In fact, each sample has only one action label. Additionally, there are not videos with multiple actors. The dataset is not multi-modality as it has only videos. Lastly, the dataset can be used for active action recognition and action recognition and not for active action detection and action detection because the actions are not spatio-temporally annotated in the videos.

2.3.6 Other Dataset Versions

2.3.6.1 Online Psychology Dataset Versions

I designed the dataset versions o.3.6 (Figure 2.11) and ob.3.6 for online psychology studies. The versions o.3.6 and bo.3.6 are identical to with only one difference: The images are blurred in ob.3.6, whereas all images in o.3.6 are clear (not blurred). The “o” stands for online while “ob” stands for online and blurred. All images of the version ob.3.6 are blurred to make the recognition task more difficult for the human participants. Each online version has just above 1.5 thousand (1,512) MVVs with 24 views, totalling about 36 thousand (36,288) SVVs. There are far less videos in the online psychology versions than in the computer vision versions because of two reasons. One, while we can relatively ask computer models to classify millions of videos, it’s unrealistic asking human participants to classify so many videos online. Two, it is usually problematic dealing with big data online.

The online versions have less conditions of in some between-SVVs dimensions than computer vision version r.3.5, to reduce the number of SVVs. However, all conditions of all between-SVVs dimensions in the online versions are included in the computer vision dataset r.3.5, even if there are some additional conditions in some between-SVVs dimensions. This relatively makes the results directly comparable between computer models and humans. In fact, there are the same 7 classes of actions in the online versions as in the computer vision version. There are only the first 4 actors of r.3.5, which are in Figure 2.9. There are all 27 3D animations per class as in r.3.5. The online versions also have both mirrors as in r.3.5. There are the same viewpoint distance ρ and the same 8 angles θ. There are only 3 angles ϕ: +45° (middle-top), +90°(middle), +135° (middle-bottom). Speaking of the within-SVV dimensions, there are 10 time points (frames) for 2-second animations which make a refresh rate of 5Hz. Each frame are squares of 224x224-pixels encoded in 3 RGB colour channels.

The Mixamo parameters of the 3D animations were also randomised in these online dataset versions as in the computer vision version. I also added some noise to the size of the actor bodies in all 3D cartesian dimensions. However, in all SVVs of online versions, background colour was grey, the light colour was white, the light intensity was constant, and the additional colour of the actor surfaces was set to 0.

I computed every blurred image in ob.3.6 with a 2D convolution between the original image in o.3.6 and a 2D gaussian kernel (filter). The kernel size was 51x51 and the 2D

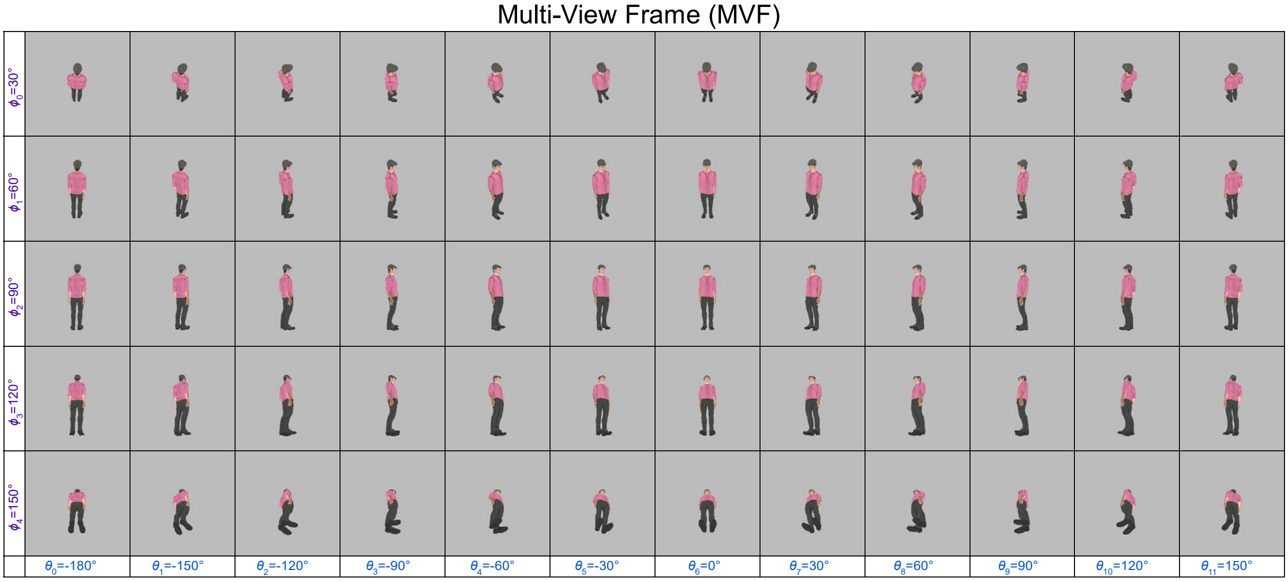


Figure 2.12. A MVF of a MVV in the pilot dataset version r.1.

gaussian σ was set to 10. I processed all these convolutions with 2 functions of the efficient OpenCV library (Bradski, 2000). I made the 2D gaussian kernel with the function cv.getGaussianKernel() and then I convolved the 2D gaussian kernel with each image of o.3.6 with the function cv.filter2D() (https://docs.opencv.org/4.x/d4/d13/tutorial_py_filtering.html). All original images of o.3.6 are stored in a GitHub repository (https://github.com/ccalafiore/dataset_o.3.6) and all blurred images of ob.3.6 in another GitHub repository (https://github.com/ccalafiore/dataset_ob.3.6).

2.3.6.2 Pilot Dataset Version

The dataset version r.1 (Figure 2.12) was an older version that I used for the first pilot studies. It contains 660 MVVs with 60 Viewpoints and so there are 39,600 SVVs. The conditions of the between-SVVs dimensions are: 6 classes of actions, 11 actors, only 5 3D animations per action class, 2 mirrors and 1 distance ρ, 12 angles θ, 5 angles ϕ. The 6 classes of actions are: pointing, discussing, waving, falling down, sitting down (standing-to-sitting), standing up (sitting-to-standing). The angles θ are: −180°, −150°, −120°, −90°, −60°, −30°, 0°, +30°, +60°, +90°, +120°, +150°. The angles ϕ are: 30°, 60°, 90°, 120°, 150. Looking at the within-SVVs dimensions, the number of frames is different for each 3D animation, but the refresh rate of all videos is at 30Hz. The numbers of frames of the videos ranged between 25 frames (0.83 seconds) and 100 frames (3.33 seconds) (*M* = 68.67, *SD* = 20.82). Finally, the size of each image is 3x256x256 (i.e., RGB colour).

In this dataset version, the only noisy parameters are the positions and the rotations of the viewpoints. As result of that, the actor hips may not be exactly at the centre of the frames as well as the size of the actors may vary on the frames of different MVVs. These noisy parameters are different for the frames of different MVVs, but they are constant for all frames of the same MVV. I did not randomize any other parameter. I set the default Mixamo parameters of the 3D animations. I did not add any noise to the size of the actor bodies, even if the viewpoint distance varies between MVVs. The background colour was light grey, the light colour was white, the light intensity was constant, and the additional colour of the actor surfaces was set to 0.

2.4 Conclusions

The aim of my thesis is inspiring researchers in robotic and human vision to study active action recognition as much as they have studied the passive action recognition. The aim of this specific chapter was providing to the scientific community the best dataset for active action recognition. This is MVVHA. To highlight why MVVHA is better choice rather than the most popular action datasets, I described strengths and weakness of MVVHA and its competitors. What stands out from these descriptions is that the current action datasets have been great for action recognition, but they are not suitable for active action recognition. In fact, they do not provide proper dynamic visual environments which change based on the actions which the active observers do in them. In other words, they do not provide different viewpoints to the observers and so the observers cannot choose position where to look at the action from. However, MVVHA does have these options because it contains MVVs, i.e. videos of the same action taken from numerous different viewpoints. The observer can decide where to look at the action from, by selecting one of the optional viewpoints.

There are a few multi-view datasets (A.-A. Liu et al., 2016; Liu et al., 2019; Shahroudy et al., 2016) which has some different viewpoints and so different options, but either their number of views is very small, their samples per classes are not numerous enough, or they do not contain many action classes. If they have not enough views, they cannot simulate the real world of an observer given that we can assume that an observer in the real world may have infinite optional views. MVVHA has many views. Furthermore, it is easy to change the number of views in the rendering code and produce different versions with as many views as a specific researcher needs for their specific study. If there are not a lot of samples per class in a dataset, they may be not enough to train the DL models. There is a huge number of samples in MVVHA. The datasets that have a little of classes, it is not representative of all possible human actions in the real word. Additionally, a few classes may make the recognition task too easy for the computer models. MVVHA has also few classes, but I will expand the number of classes in the future. To my best knowledge, MVVHA is currently the best dataset that researchers can use for their active action recognition studies, even if it still presents several limitations.

Of course, MVVHA has its own weakness and I plan to overcome those. As I have already mentioned, I will extend the number of classes from 7 to hundreds. MVVHA has a little of data variability with the randomization of background colour, light colour and light intensity. I will increase the data variability by adding many random scene backgrounds with random objects. I can do videos with several actors who do different action and have different action labels. Therefore, I will also give multiple labels for each video. I can add multi-modality data such as the 3D location of the actor skeletons and the depth maps. Finally, I will also add some time and space annotations of the action in the videos. Coping with these limitations will make an even better dataset than MVVHA for active action recognition and detection. Now, the MVVHA is the best solution for active vision studies and researchers can use it for their early studies of this young field.

3 Piloting Active Action Recognition of Human Observers

3.1 Introduction

For a safe and productive interaction between agents such as humans and robots, it is crucial that these agents recognize each other's actions, such that they can provide the most appropriate reaction to any social event. For instance, by recognizing some particular actions of a person, they can infer this person is injured and give them assistance. They also need to recognize a criminal action of a person to report them to the police. Humans can easily recognize the actions of an actor. However, in real life, human observers do not recognize actions passively. Instead, they select the positions from where to look at the actions clearly. Yet, most studies about human vision have examined the passive vision where the viewpoint was predetermined for the observers. This study aims to investigate the active action classification of humans. Specifically, it aims to reveal whether and how people select views efficiently for more accurate and faster action classification. This new knowledge can inspire the field of robotic vision to design active robots that seek important information (e.g., injured people or criminals).

Only a few studies have been focused on how the position of the observation can influence action recognition of humans and which viewpoints humans tend to select more often for more accurate and faster action recognition. Mitchell and Curry (2016) looked at how people recognise themselves and the others in point-light displays (PLDs) of their walks from 2 different views: frontal view and half-profile view. The recognition of the self and others in PLD walks from the frontal view was as accurate as from the half-profile view. However, they studied the recognition of actors rather than the actions, they used only two views, and these views were forced to the participants.

Because of natural selection, humans should choose their viewpoints efficiently to understand the surrounding environment. In fact, in the wild, labelling the surrounding environment as either dangerous or safe as rapid and accurate as possible was essential to survive. The efficient choice of the viewpoints made environment classification quicker and more accurate. Thus, efficient active action recognition improved the likelihood of people to survive and should have been naturally selected.

There were two main objectives in this pilot study. The first one was to examine whether human participants have higher performance recognizing actions when they are allowed to select the position from where to look at the actions. I measured the performance of action recognition with both accuracy and RT. The higher performance conditions were the experimental conditions where the participants scored higher accuracy or shorter RT, while the conditions with lower performance were the ones in which the accuracy of the participants was lower or their RT was longer. If people have the efficient skill of looking around for social information, the action recognition of my participants should be more accurate or faster when they have the opportunity to select the efficient views than when they cannot choose their viewpoints.

The second main objective was to inspect whether human participants select the efficient views more often than the inefficient views during active action recognition. The efficient views are the views with which participants scored higher action recognition performance (higher accuracy or shorter RT) while the inefficient viewpoints are the ones with which participants had lower performance (lower accuracy or longer RT) in recognizing actions. If the participants efficiently choose the position from where to look at an action, they should select the efficient views more often than the inefficient views.

There were two other subordinate objectives in this study to achieve the second main objective. The first one was to determine whether the action recognition with the MVVHAs is viewpoint-dependent, meaning that the action recognition performance is affected by changes in viewpoint. The difference between efficient and inefficient viewpoints is only valid if changes in viewpoint significantly affect the recognition performance. If it were viewpoint-dependent, the second subordinate objective was to detect the efficient and inefficient viewpoints for action recognition. After achieving the second subordinate objective and knowing the efficient and inefficient views, I could accomplish the second main objective by examining whether the human participants choose the efficient views more often than the inefficient views during active action recognition.

To achieve these objectives, I stimulated some participants with the pilot version of my own dataset MVVHA which has 60 views and asked them to actively classify the actions in each MVV. Within a trial, every participant can only see one of the 60 SVVs at the time. However, their view could change 3 times in each trial simulating a movement of the participant viewpoint in the 3D space. I manipulated within-subjects the type of viewpoint movement which had two trial conditions: Random Movement (RM) and Self-Controlled Movement (SCM). In both RM and SCM trials, participants had to classify an action of an actor displayed on the screen. However, in the SCM trials, the participants could choose their viewpoints before classifying the action, while, in the RM trials, their viewpoint changed randomly. In this experimental setup, the task of the participant in each trial was to move their view 3 times and then classify the action of the actor on the screen.

There were three hypotheses in this study. The first one was the following: the action recognition performance of the participants would be higher in the SCM condition than in the RM condition. That is because the task of participants in the SCM condition was active action recognition and they can efficiently select the viewpoints, while the task in the RM condition was passive action recognition and cannot choose the efficient viewpoints.

The second hypothesis was that the action recognition in MVVHAs is viewpoint-dependent because the action classes in MVVHA are similar to each other. For example, discussing, pointing and waving are very similar, dancing and falling down are slightly similar, and sitting down and standing up are practically the same if we exclude the time dimensionality. This relation between similarities of object shapes and the viewpoint dependency of object recognition performance was unanimously found in several studies (Hayward & Williams, 2000; Leek & Johnston, 2006; Tarr & Hayward, 2017; Tarr & Pinker, 1990; Tjan, 2001). See the introduction in chapter 4 for an extensive literature review of viewpoint-dependent and viewpoint-invariant object recognition performances. Even though I aimed to highlight the efficient and inefficient viewpoints, I did not make any predictions about which viewpoints were efficient or inefficient.

The third hypothesis was that participants would select more often the efficient views than the inefficient views in the SCM trials. The participants in the SCM condition did active action recognition. Thus, if they can efficiently choose the position from where to look at the actions, they should select the efficient views more often than chance and choose the inefficient views less often than chance.

3.2 Methods

3.2.1 Participants

I tested 13 participants. They were students of the University of Essex. 31% (4) of them were male whereas the rest 69% (9) were female. Their age ranged from 19 to 22 (*M* =

19.77, *SD* = 1.17). They were rewarded with money or academic credits.

3.2.2 Materials

As stimuli, I used the pilot version of my own dataset MVVHA (see chapter 2). I used 360 MVVs out 660 from this dataset. Each of them has 60 SVVs of same actor performing the same action from 60 different viewpoints. Thus, I only took 21,600 SVVs out 39,600. From this version of the dataset, I selected 6 classes of actions, only 6 (out of 11) actors, 5 3D animations per class of actions, 12 angles θ and 5 angles ϕ. There were 30 VR animations in total (6 classes x 5 animations per class). The 6 classes of actions were: pointing, discussing, waving, falling down, sitting down (standing-to-sitting), standing up (sitting-to-standing). Every viewpoint can either have one of the following angles ϕ:

$$\phi \in [30°, 60°, 90°, 120°, 150°]$$

while its angle θ can either be one of the 12 angles below:

$$\theta \in [-180°, -150°, -120°, -90°, -60°, -30°, 0°, 30°, 60°, 90°, 120°, 150°]$$

The image size of each single-view video is 256x256x3 (i.e., RGB colour). Videos were recorded at 30Hz and the number of frames (duration) was different for each 3D animation. The numbers of frames of the videos ranged between 25 frames (0.83 seconds) and 100 frames (3.33 seconds) (*M* = 68.67, *SD* = 20.82).

3.2.3 Design

I manipulated three independent variables within-subject: type of viewpoint movements; angle ϕ of the starting viewpoints; angle θ of the view a trial starts with. There were two types of view movements: RM and SCM. In the RM trials, the views changed randomly, while the in SCM ones, the participants chose the within-trial movements of the viewpoints.

I chose 5 conditions of the elevation angle ϕ of the starting view. So, a trial started by showing the action from a viewpoint that could either have ϕ=30° (top views), ϕ=60° (middle-top views), ϕ=90° (middle views), ϕ=120° (middle-bottom views) or ϕ=150° (bottom views). The third independent variable was the angle θ of the starting view. The angle θ rotates round the z-axis or the actor. I used 12 conditions of beginning angle θ. A trial started with a view whose angle θ could either be: −180° (back views), −150° (right-back views), −120° (right-back views), −90° (right views), −60° (right-front views), −30° (right-front views), 0° (front views), 30° (left-front views), 60° (left-front views), 90° (left views), 120° (left-back views), 150° (left-back views).

In addition, there were three dependent variables: accuracy of action recognitions; RTs of action classification; percentage frequencies of the ending views. I looked at whether the three independent variables influence accuracy and RT of action classification. Additionally, the percentage frequency of a view at the end of the trials could tell us how often the participants selected that particular view compared to the other views. I expected that the participants would choose the efficient views more often the inefficient views. The efficient views are the ones that produced higher accuracy and lower RT of action classification as starting viewpoints.

3.2.4 Procedure

I tested every participant within a session of approximately 1 hour and 30 minutes. In the session, each participant saw 720 trials. These 720 trials were constructed by repeating each of the 360 MVVs twice: one time per each of the two conditions of type of viewpoint movements. The two conditions of the type of view movements were RM and SCM. Therefore, in total, I had 360 RM trials and 360 SCM trials. In the RM trials, the position of the view changed randomly whereas in the SCM trials, the view movements were controlled by the participants.

I split the 720 trials in 6 blocks of 120 trials. In each block, there were 60 RM trials and 60 SCM trials. The order of the RM and SCM trials was pseudo-randomized within each block. The 1st block was used for familiarization and there were then 5 blocks for the main experiment. I did not analyse the 120 trials of the familiarization. I only analysed the 600 trials of the main experiment. The aim of the familiarization was to give the participant the chance to practice with the lab task and make sure the participants understood the task before doing the actual test. Between blocks, participants had short breaks of a few minutes to reduce participants' fatigue.

I showed to each participant the 60 MVVs of one actor twice in the block of familiarization. This actor was always the actor_00. Every participant saw all remaining 300 MMVs of the other 5 actors twice in the 5 blocks of the main experiment. The actors of the main experiment were always actor_01, actor_02, actor_03, actor_04 and actor_05 for each participant. In this way, the analyses came from the performances given the same 5 actors of the main experiment as stimuli.

As the Figure 3.1 shows, both RM and SCM trials start by showing a SVV of a MVV from a random starting viewpoint. The starting viewpoint was one of the 60 views (or one SVV) of each MVV. The 60 views are defined by the 60 combinations of the 5 conditions of the angle ϕ and the 12 possible conditions of the angle θ. For each participant, I randomly split the 300 MVVs of the main experiment in 60 groups of 5 MVVs. I randomly assigned a specific view to one of the 60 groups of 5 MVVs and then, for each of the 60 groups of 5 MVVs, I made 5 RM trials and 5 SCM trials with a different starting view. Note that for a given participant, the starting viewpoint was the same one in both the RM trial and the SCM trial with the same MVV.

Following the start of the trial, the participants could move their own view to any of the 8 neighbouring views with respect to the current view (i.e., the top, bottom, left, right, top-left, top-right, bottom-left, or bottom-right neighbouring view) by pressing a button. In the SCM trials, after they made a movement the SVV continued from the viewpoint they selected. For instance, if they selected to move the view to the top-left neighbouring view respect to

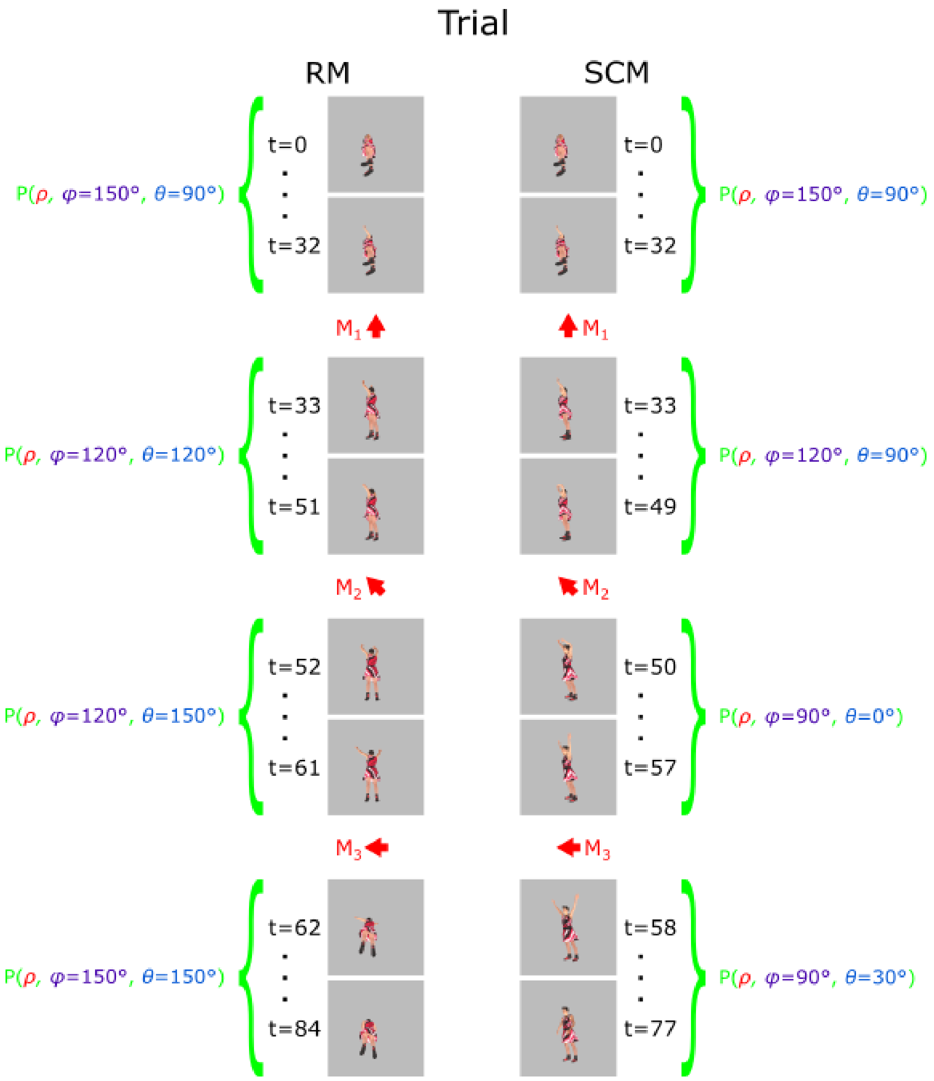


Figure 3.1. A potential RM trial (left) and a potential SCM trial (right). Both start by playing the same video from the frames (timepoints) t=0 to t=32 with the view (ϕ=150°,θ=90°). Then, the participant made the first view movement M_1 to the upper view at the timepoint t=32 and the video continued with another view from the frames t=33 to t=51 in RM and from the frames t=33 to t=49 SCM trials. The view actually switched to the upper view (ϕ=120°,θ=90°) in the SCM trial, while it randomly went to the right-upper view (ϕ=120°,θ=120°) in the RM trial. Next, the participant moved to the left-upper view for M_2 in both RM and SCM trials and to the left view for M_3 in both RM and SCM trials. The view actually moved to the direction requested by the participant in the SCM trial and to a random direction in the RM trial even for the moves M_2 and M_3.

the current one, they kept watching from that chosen view. The RM trial was exactly the same. The only difference was that in the RM trials, when the participant selected one of the 8 possible neighbouring views, the view moved to a random neighbouring view. Note that even if in the RM condition the viewpoint change randomly, participants still had to select by button press one of the 8 neighbouring view to move like in the SCM condition. The only difference between the two conditions of viewpoint movement is that when they pressed one of the 8 buttons corresponding to the 8 possible viewpoint movements, the viewpoint changed randomly in the RM condition and it changed according to the participant selection in the SCM condition. The participant could never spot that they were in a RM trial or in SCM trial before they made the first movement and notice the type of view movement.

In both RM trials and SCM trials, they could only move 3 times and they had to move at least 3 times before classifying the action performed by the actor. Hence, the number of movements was always 3 in any trial. Note that when they changed their own view, the video continued to play smoothly. The frames were displayed at 30Hz to reproduce the same action speed of the original 3D animations.

After the 3 movements, both the RM and SCM trials either ended at the action classification of the participant or after 30 seconds from the start of the trial if the participant does not classify. The participants classified the action by pressing one of the 6 keys to identify which action the actor was performing. The classification was thus a 6-alternative forced choice task. The 6 possible action classes were pointing, discussing, waving, falling, standing-to-sitting and sitting-to-standing. Hence, they were asked to select one of those 6 options in each trial based on which action they recognized in that specific trial. They had to press 1 for pointing, 2 for discussing, 3 for waving, 4 for falling, 5 for standing-to-sitting and 6 for sitting-to-standing. The options were displayed on the bottom of the screen throughout the trial. If the participant does not classify the action in a trial, the short video of that trial was replayed up to 30 seconds and then the stimulation script moved on to the next trial.

Between trials there was an inter-trial interval of 1 second plus a jitter of 0.5 seconds (i.e a random time interval between 0 and 0.5 seconds). At this point, a fixation cross was displayed in the centre of the screen along the 6 options corresponding to the 6 classes of actions. The location of the fixation cross or the centre of the screen was also the location of the hips of the actor during the trial.

The experimenter told to the participants the instructions below at the beginning of the test.

> *"I will play some short videos of a few seconds on this screen. You are asked to watch these videos and classify the actions that you see in them. There are six possible types of actions in the videos: pointing, discussing, waving, falling down, standing up and sitting down. The videos will be displayed one at a time. There is only one person in each video that does only one action.*
>
> *Once a video is played, please press one of six specific keys on the keyboard as quickly and accurately as possible to indicate the type of action in each video. Press 1 for pointing, 2 for discussing, 3 for waving, 4 for falling down, 5 for standing up and 6 for sitting down. Do not worry if you do not remember these matches of numbers and actions at this point. They will always be displayed on the bottom of the screen over the whole test to remind you.*
>
> *Each video will be played and replayed for a maximum of 30 seconds until you press one of the six keys. Therefore, you have a maximum of 30*

seconds to indicate the action of any video. If you do not respond for 30 seconds since the start of a video, we will move on, and the next video will be played.

Additionally, while a video is played, you have to move your viewpoint where you look at the action towards another good viewpoint. A good viewpoint is a viewpoint from where the action is more clear and more recognisable than from most other viewpoints. You have to make a minimum of three and a maximum of three view movements per video. You can move your viewpoint to left, top, right, and bottom, as well as top-left, top-right, bottom-left, and bottom-right, by pressing one or two of the four arrow keys on the keyboard. However, half of times, your viewpoint will actually move in the direction you chose, while the other half of times, your view will move in a random direction.

In short, when a video starts, you have to move your viewpoints three times towards another good viewpoint, and then classify the action as quickly and accurately as possible.

The videos are split into six blocks. Each block will last about 8 minutes. You will have the opportunity to rest for a few minutes between the blocks. The first block of videos is the familiarization, and the other five blocks are the real experiment. I will not analyse the data of the familiarization. The familiarization is only for you to learn and practice the task. In the familiarization, once you classify the action of a video, the screen will display feedback about your classification. The feedback will say correct if your classification was correct, or incorrect if it was wrong. In the real *experiment, no feedback will be displayed.”*

3.3 Results

3.3.1 The Efficient Views

I defined efficient views as the views that generated higher performance in the action recognition task of my participants. In practice, higher performance views were the ones with higher accuracy and shorter RTs. Inefficient views were those which led to lower accuracy and longer RTs.

A three-way ANOVA was conducted to highlight the main effects of the type of viewpoint movements, of the angle ϕ and of the angle θ of the starting view, and their interaction effects on accuracy and RT. All effects are reported as significant at $p < .05$.

In general, accuracy was very high with participants performing accurately throughout. There was a significant main effect of angle ϕ on accuracy, $F(4, 48) = 3.68$, $p = .011$. The elevation of the starting viewpoint did have an impact on the accuracy of action recognition. To follow-up this effect, I ran multiple pairwise comparisons in accuracy between the five conditions of angle ϕ by using t-tests with Bonferroni correction. Comparing the five elevation levels would have required ten comparisons, and these many comparisons would have led to a large probability of type I errors. Therefore, the p-values of the multiple pairwise comparisons were corrected by the Bonferroni method. Although the top elevation angle ϕ_0=30° ($M = .972$, $SD = .030$) was nearly more accurate than the bottom elevation angle ϕ_4=150° ($M = .944$, $SD = .036$), $t(12) = -3.29$, $p = .065$, all ten possible pairwise comparisons of the five conditions showed no significant difference in accuracy. This may have been because the Bonferroni correction is too conservative for many comparisons.

Figure 3.2(A) highlights the general pattern of the action recognition accuracy in the five different conditions of angle ϕ. The accuracy as a function of the elevation angle ϕ steadily decreased from top to bottom viewpoints. The top viewpoints with elevation angle ϕ_0=30° (*M* = .972, *SD* = .030) were the most accurate. The accuracy slightly decreased at the next middle-top viewpoints with angle ϕ_1=60° (*M* = .963, *SD* = .032) and remained flat in the middle ϕ_2=90° (*M* = .967, *SD* = 0.014) and the middle-bottom viewpoints ϕ_3=120° (*M* = .962, *SD* = 0.023). Finally, it dropped in the bottom viewpoints ϕ_4=150° (*M* = .944, *SD* = .036) which were the least accurate. Overall, participants' accuracy was lower when they looked at the actions from the bottom views and was higher from the other views, ranging from the middle-bottom to the top views.

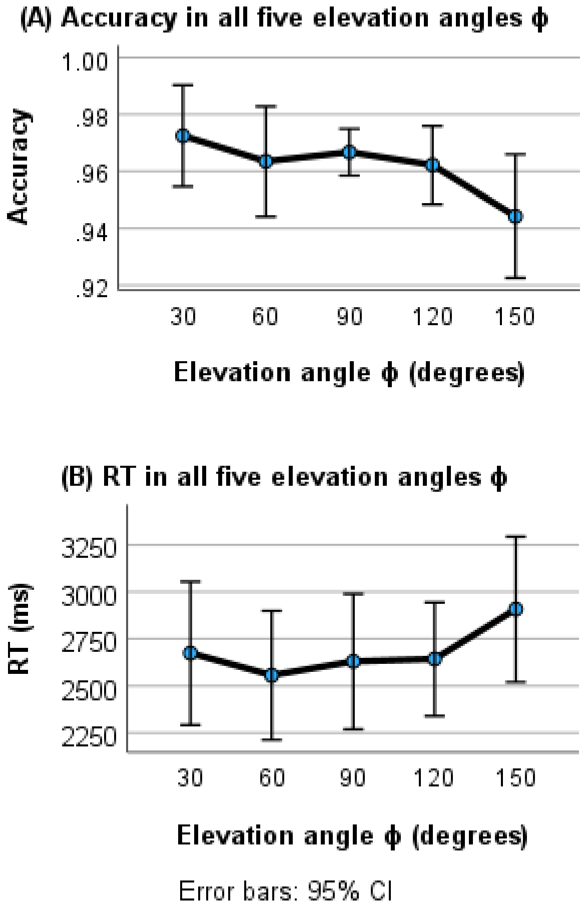
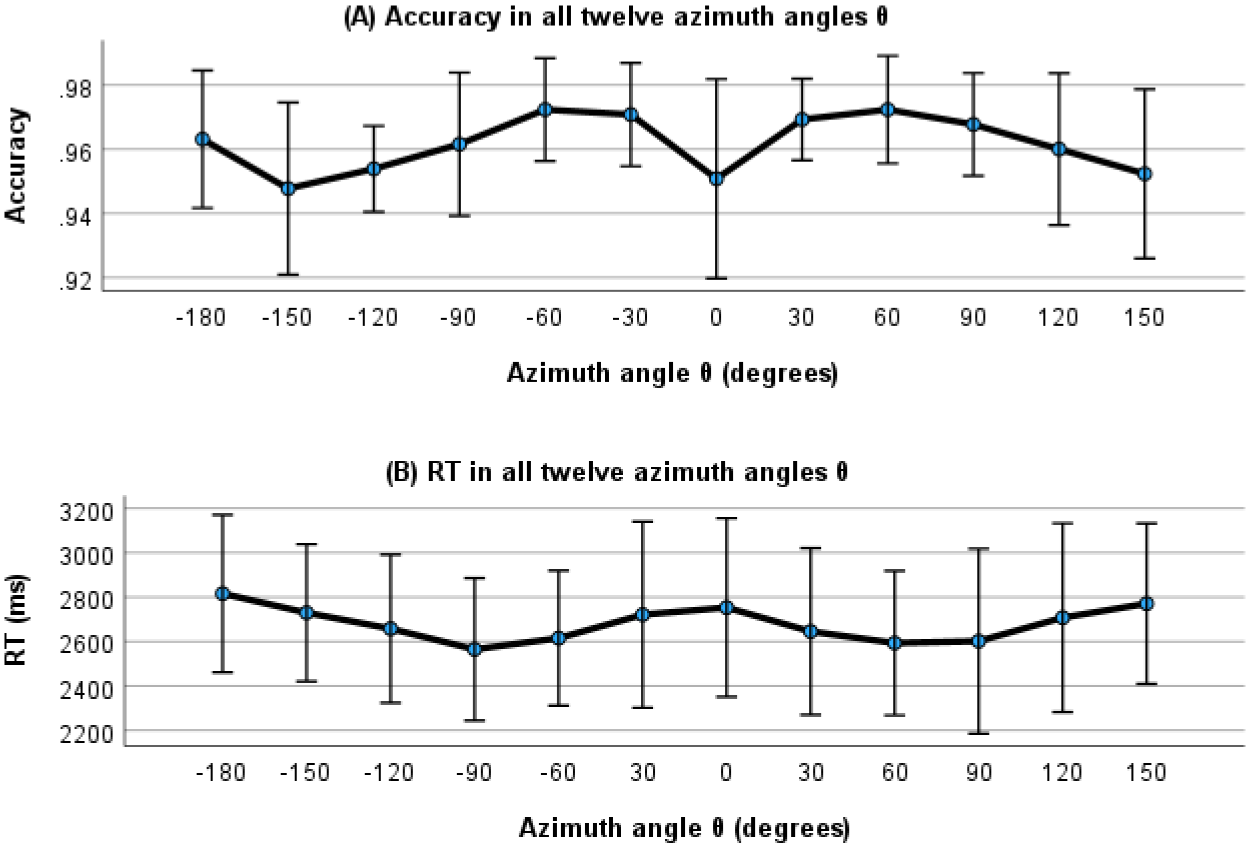


Figure 3.2. The action classification accuracies (A) and the RTs (B) given each angle ϕ of the starting view, regardless of the angle θ of the starting view and the type of view movements.

Furthermore, there was also a significant main effect of angle ϕ on RT, *F*(4, 48) = 17.52, *p* < .001. This suggests that view elevation influenced the speed of action recognition of the participants. To further investigate this main effect on RT, multiple pairwise comparisons were performed between the RTs of the 5 conditions of the angle ϕ. The p-values were corrected with the conservative Bonferroni method. RT in the bottom views with ϕ_4=150°

(*M* = 2,907 ms, *SD* = 641 ms) was longer than in the top views ϕ_0=30° (*M* = 2,673 ms, *SD* = 631 ms), *t*(12) = 4.16, *p* = .013, middle-top views ϕ_1=60° (*M* = 2,555 ms, *SD* = 568 ms), *t*(12) = 6.97, *p* < .001, middle views ϕ_2=90° (*M* = 2,629 ms, *SD* = 595 ms), *t*(12) = 6.74, *p* < .001, and middle-bottom views ϕ_3=120° (*M* = 2,641 ms, *SD* = 500 ms), *t*(12) = 4.22, *p* = .012. RT in the middle-top views ϕ_1=60° (*M* = 2,555 ms, *SD* = 568 ms) was significantly shorter than in the top views ϕ_0=30° (*M* = 2,673 ms, *SD* = 631 ms), *t*(12) = −3.74, *p* = .028, and the bottom viewpoints ϕ_4=150° (*M* = 2,907 ms, *SD* = 641 ms), *t*(12) = −6.97, *p* < .001. However, the middle-top elevation ϕ_1=60° (*M* = 2,555 ms, *SD* = 568 ms) was not significantly different from the middle elevation angle ϕ_2=90° (*M* = 2,629 ms, *SD* = 595 ms), *t*(12) = −2.27, *p* = .421, and the middle-bottom angle ϕ_3=120° (*M* = 2,641 ms, *SD* = 500 ms), *t*(12) = −2.82, *p* = .154. Figure 3.2(B) displays the RTs of action classification in each of the 5 levels of the angle ϕ. Overall, action recognition was faster when participants watched the actors from the middle-top views and was slower when they looked from the bottom viewpoints.

There was not a significant main effect of the angle θ on accuracy, *F*(4.42, 52.03) = 1.26, *p* = .295. However, Figure 3.3(A) shows a marginal pattern of accuracy in the viewpoint angles θ. First, accuracy was generally lower in the back views θ_1=−150° (right-back), θ_2=−120° (right-back), θ_{10}=+120° (right-back), θ_{11}=+150° (left-back), and higher in the front-side views θ_4=−60° (right-front), θ_5=−30° (right-front), θ_7=+30° (left-front) and θ_8=+60° (left-front). Surprisingly, there was a drop in accuracy at the front views θ_6=0°. Second, the accuracy in the right and left views was symmetric, meaning that the right-back, right and right-front views were equivalent to the left-back, left and left front, respectively. To further investigate this marginal pattern, multiple pairwise comparisons were performed with the Bonferroni correction. All 66 comparisons showed no significant difference between the twelve viewpoint angles θ. Corrections for multiple comparisons were extremely



Error bars: 95% CI

Figure 3.3. The action classification accuracies (A) and the RTs (B) given each angle θ of the starting view, regardless of the angle ϕ of the starting view and the type of view movements.

conservative with such a large number of comparisons.

There was a significant main effect of the angle θ in RT, *F*(11, 132) = 2.72, *p* = .003. Figure 3.3(B) highlights the pattern of RT as function of the angle θ. This is opposite to the one on accuracy. Similarly to the accuracy, RT in the right and left view angles was symmetric. The RT peaks in the back view θ_0=−180° and then, as we move to either left or right sides of the actor, it decreases significantly reaching its lows in the right views θ_3=−90°, in the right-front views θ_4=−60°, in the left-front viewpoints θ_8=+60° and left views θ_9=+90°. Subsequently, it increases as we move to the front, peaking again in the front views with θ_6=0°. In other words, action recognition RT was long in the back, right-back,

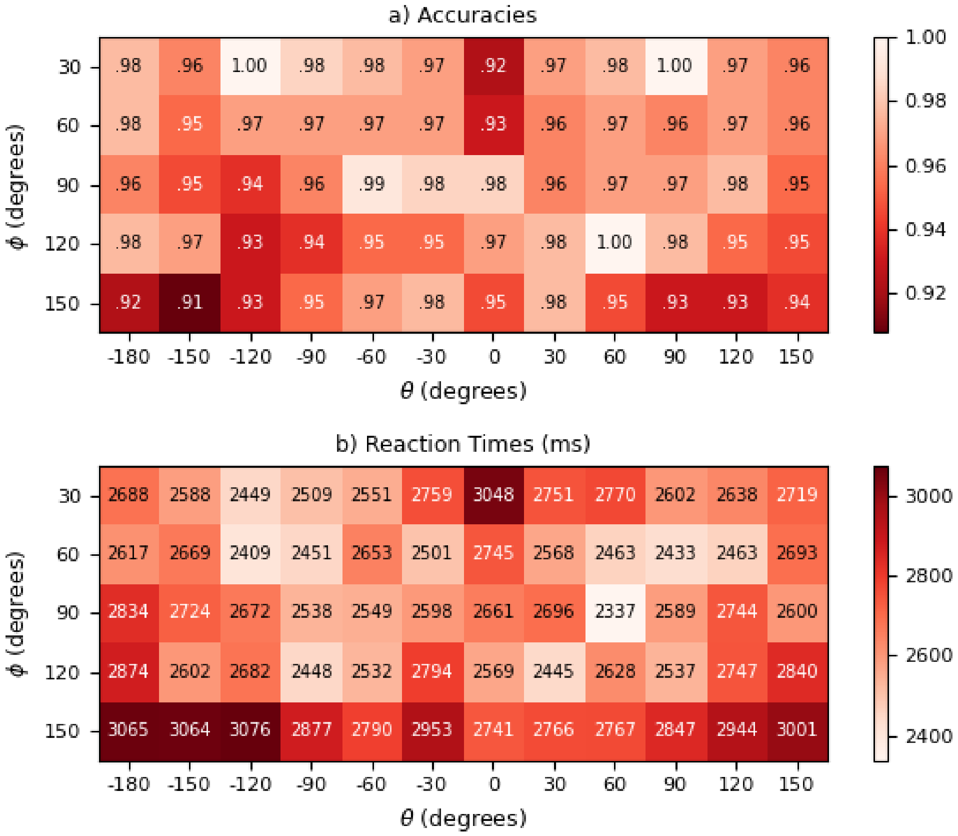


Figure 3.4. The action classification accuracies (a) and the RTs (b) given each combination of the two angles ϕ and θ of the starting view, regardless of the type of view movements.

left-back and front views $\theta_0=-180°$, $\theta_1=-150°$, $\theta_{11}=+150°$, $\theta_6=0°$, while it was short in the right, right-front, left-front and left views $\theta_3=-90°$, $\theta_4=-60°$, $\theta_8=+60°$, $\theta_9=+90°$. To investigate this effect, I ran multiple pairwise comparisons with the Bonferroni correction. There was no significant difference in all 66 comparisons of the twelve conditions. This is again because the Bonferroni correction is enormously conservative when the comparisons are so many.

Although in accuracy there was no significant interaction between the elevation angle ϕ and the azimuth angle θ of the starting viewpoints, $F(44, 528) = 1.07$, $p = .360$, the interaction of the same variables in RT was significant, $F(44, 528) = 1.48$, $p = .026$. Therefore, while the effect of the angle θ on accuracy does not depend on ϕ, it does on RT. The two heatmaps in Figure 3.4 visualize the interaction between the angles ϕ and θ of the starting viewpoints on accuracy and RT.

3.3.2 Performance of Action Recognition is Better in the Active Condition

If participants select their own views efficiently, they should get higher performance in the SCM condition than in RM condition. That is because in the SCM condition they could select efficient views while they cannot do so in the RM condition. This difference in performance should be particularly clear when the trial starts with an inefficient view as in the SCM condition participants should be able to go from an inefficient view to an efficient view. In other words, there should be an interaction effect between the viewpoints and the type of movement. Specifically, we should get expect an interaction effect in accuracy and RT between the angle ϕ of the starting viewpoint and the viewpoint movement type, between the angle θ of the starting viewpoint and the viewpoint movement type and ideally 3-way interaction between the angle ϕ of the starting viewpoint, the angle θ of the starting viewpoint and the type of viewpoint moments.

In accuracy, there was not a significant

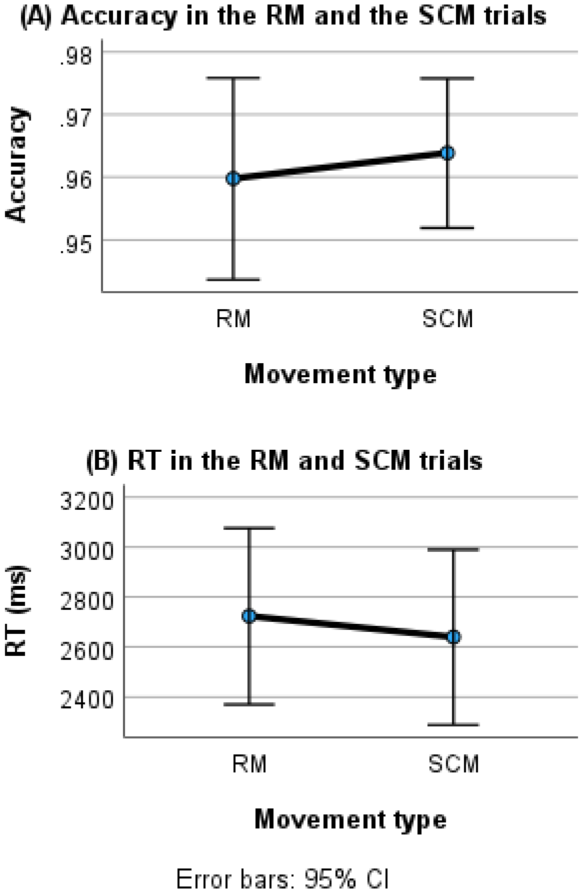
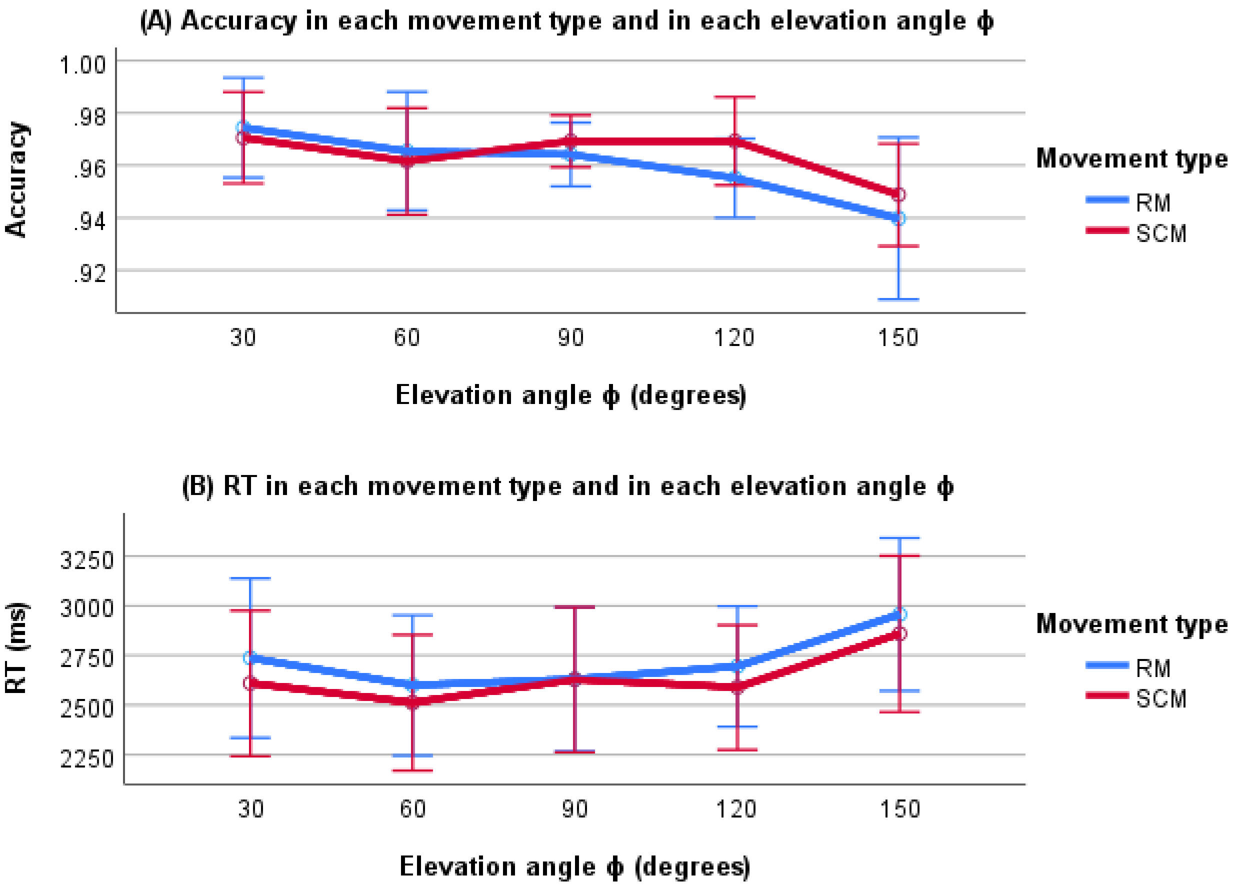


Figure 3.5. The action classification accuracies (A) and the RTs (B) in each type of view movements, regardless of both angles ϕ and θ of the starting view.


(A) Accuracy in each movement type and in each elevation angle ϕ

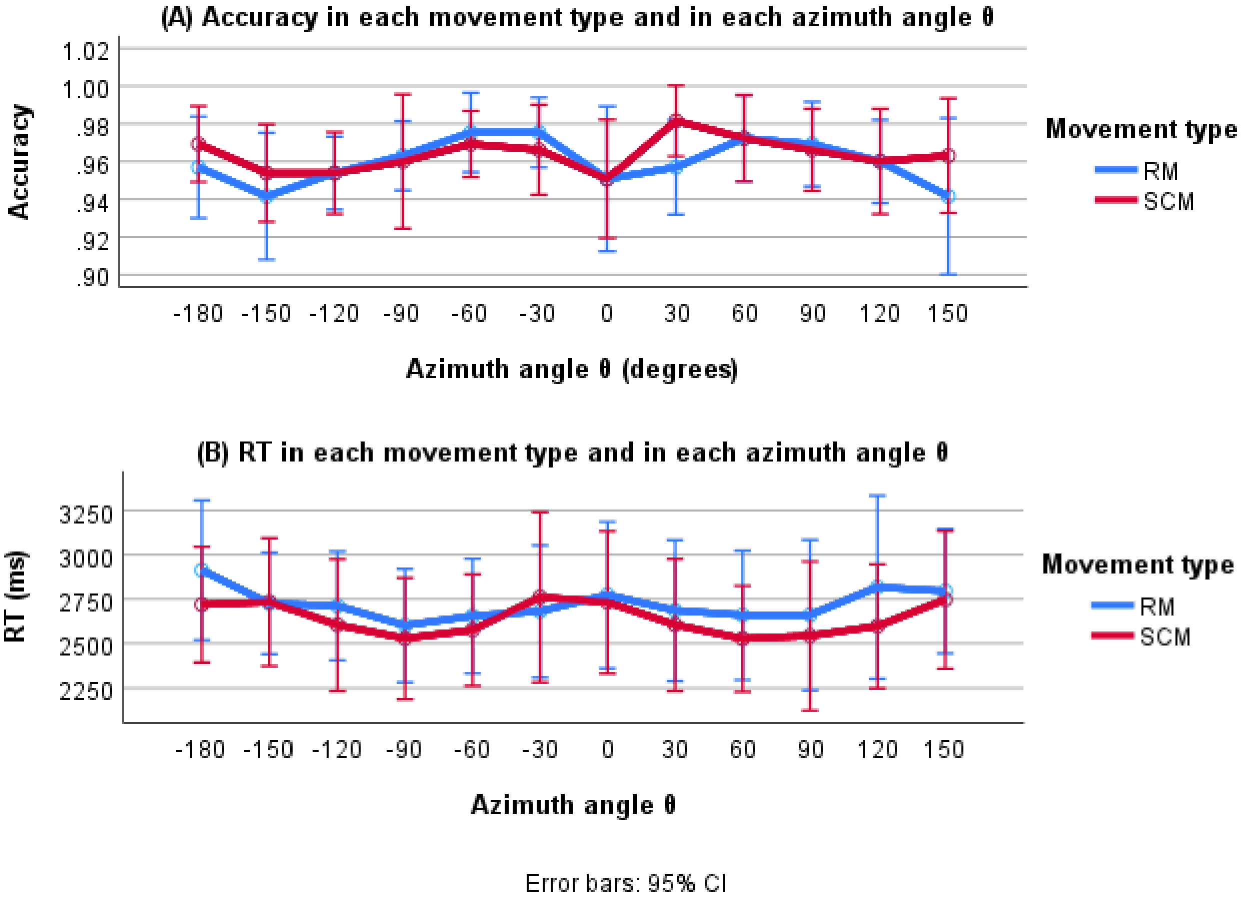
(B) RT in each movement type and in each elevation angle ϕ

Error bars: 95% CI

Figure 3.6. The action classification accuracies (A) and the RTs (B) given each combination of the type of view movements and the angle ϕ of the starting view, regardless of the angle θ.

main effect of type of viewpoint movement type, *F*(1, 12) = .63, *p* = .445, even if the action recognition in the SCM condition (*M* = .964, *SD* = .020) was slightly more accurate than in the RM condition (*M* = .960, *SD* = .027). Figure 3.5(A) displays the accuracies of the two movement conditions. However, my analysis revealed a significant main effect of view movement type on RT, *F*(1, 12) = 12.79, *p* = .004. This suggests that the action recognition of the participants was faster in the SCM condition (*M* = 2,639 ms, *SD* = 580 ms) than in the RM condition (*M* = 2,723 ms, *SD* = 584 ms). Figure 3.5(B) highlights this pattern.

The interaction effect on accuracy between movement type and angle ϕ of the starting


(A) Accuracy in each movement type and in each azimuth angle θ

(B) RT in each movement type and in each azimuth angle θ

Figure 3.7. The action classification accuracies (A) and the RTs (B) given each combination of the type of view movements and the angle θ of the starting view, regardless of the angle ϕ.

views was not statistically significant, *F*(4, 48) = 1.11, *p* = .361. The effect of the starting angle ϕ on accuracy was not different in the two different view movement types. In Figure 3.6(A), we can see the interaction plot of accuracy between movement type and angle ϕ. There was not a significant interaction effect on RT between movement type and angle ϕ of the starting views, *F*(4, 48) = 1.15, *p* = .346. The effect of the starting angle ϕ on RT was not dependent on the view movement type. Figure 3.6(B) contains the interaction plot between movement type and angle ϕ for RT.

The interaction effect between movement type and angle θ of the starting views was not significant on both accuracy *F*(11, 132) = .61, *p* = .814, and RT, *F*(11, 132) = 1.19, *p* =

.298. I ran twelve paired t-tests with the Bonferroni correction to compare the RTs in the RM condition and in the SCM condition for each of the twelve horizontal angles θ. There was not any significant difference between the two movement conditions in every condition of the angle θ. The two interaction plots between movement type and angle θ is in Figure 3.7. The effect of the angle θ of the starting viewpoint on both accuracy and RT did not depends on the viewpoint movement type.

Finally, there was no significant 3-way integration effect between movement type, angle ϕ and angle θ of the starting views on both accuracy, *F*(44, 528) = .97, *p* = .537, and RT, *F*(44, 528) = 1.03, *p* = .419.

3.3.3 Efficient views are selected more often

Let us now turn to the main question about where participants move their own viewpoints. The results above showing the effects of the starting views on the accuracy and RT of action recognition highlighted the efficient and the inefficient views. The efficient views are the ones which participants had higher accuracy and lower RT with. Those tuned to be the upper and front-sided views. The inefficient views were the ones that as starting viewpoints produced lower accuracy and higher RT of action recognition. These were the back and bottom views, which is not surprising given the stimuli involved and our experience of observing people from the front vs. the back. If participants select their own viewpoint efficiently, they should select the efficient views more often than chance and they should also select the inefficient views less often than chance. To measure how often they selected each viewpoint, I computed the percentage frequencies of the ending viewpoints. The percentage frequency of a specific ending view was the empirical percentage of trials that ended with this view. In practice, for a given view, I computed the empirical percentage of trials that ended with that specific view. I did that for all 60 views individually and obtained 60 percentage frequencies of the ending viewpoints. A viewpoint with high percentage frequency of ending views in the SCM trials means that it was chosen often by the participants. Instead, a view with low percentage frequency of the ending view is evidence that participants rarely chose that specific view. Additionally, if they chose the efficient views more often, there should be a positive correlation between the accuracy and the percentage frequency of the ending views. Furthermore, for the same reason, the RTs and the percentages of the ending views should be negatively correlated.

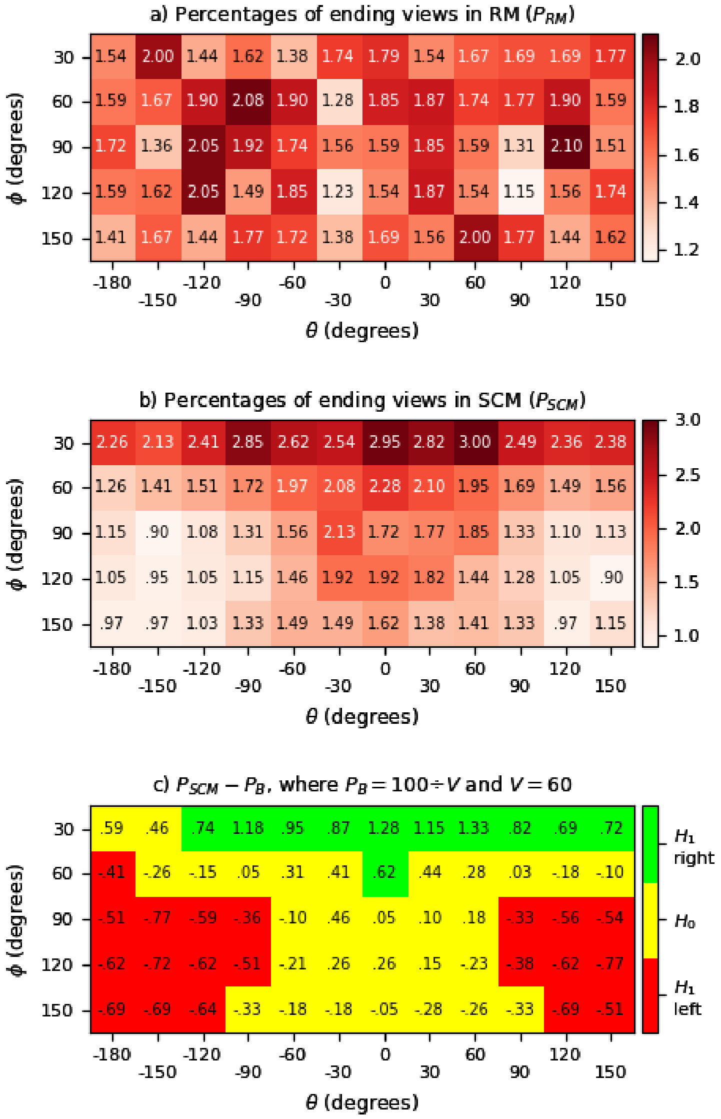


Figure 3.8. The percentage frequencies of the views at the last timepoint just before the action classification of the participants in all RM trials (a) and all SCM trials (b). (c) The difference between the percentage frequencies of the ending views in the SCM condition and the baseline 1.67 (100/60).

Figure 3.8(a) shows the percentage frequencies of ending views in the RM condition, whereas Figure 3.8(b) highlights the percentage frequencies in the SCM condition. In these two charts, each square in the heatmaps corresponds to 1 of the 60 possible views (the same views depicted in Figure 2.12). The colour of each view (square) is a representation of how high the percentage frequency of that view is compared to the percentage frequencies of the other views of the same movement condition. The number on each square is the actual percentage frequency of that specific view. There were 60 possible views, so the percentage frequency baseline for each view was 1.67 (100 / 60 = 1.67). The baseline was the theoretical percentage frequency of each viewpoint if the viewpoints were selected at random, similarly to the RM condition.

Figure 3.8(a) shows that the percentage frequencies of the ending views in the RM condition were indeed random and the percentage frequencies of all viewpoints did not considerably differ from the baseline. On the other hand, the SCM percentage frequencies of the ending views in Figure 3.8(b) have a consistent pattern. They were very low in the back and bottom views and steadily increased going towards the front and top views where they are very high.

This pattern of the percentage frequencies of the ending views in the SCM condition is also emphasised by the heatmap in Figure 3.8(c) which displays the differences between the ending view percentage frequencies in the SCM condition and the baseline. Thus, positive numbers on some views mean that the percentage frequencies of these views in the SCM condition were higher than the baseline, while negative numbers on the other views indicate that the SCM percentage frequencies of these other views were lower than the baseline. In the red views, the percentage frequencies were largely lower than the baseline. In the yellow viewpoints, these differences were moderate. Finally, in the green views, the percentage frequencies of ending views in the SCM were a lot higher than the chance. A way to interpret the colour in this heatmap is that participants moved away from the red views and towards the green views. The results in Figure 3.8(b) and Figure 3.8(c) suggest that participants tended to select the top and front views and they rarely chose the bottom and back view. Therefore, they tended to move their view away from the inefficient views towards the efficient ones in line with my predictions.

To provide more statistical evidence for the efficient selection of the views, I computed a Pearson's correlation between the accuracy and the ending percentage of all 60 views. These variables were positively correlated, *r*(58) = .35, *p* = .006. The results suggest that participants selected more often the high-accuracy views. Figure 3.9(a) shows the relations of the accuracies and ending percentage frequencies of the 60 views. In addition, the RTs and the ending percentage frequencies of the views were negatively correlated, a *r*(58) = −.44, *p* < .001. This correlation suggests that the participants selected more often the "fast views". Figure 3.9(b) highlights negative correlation of the two variables.

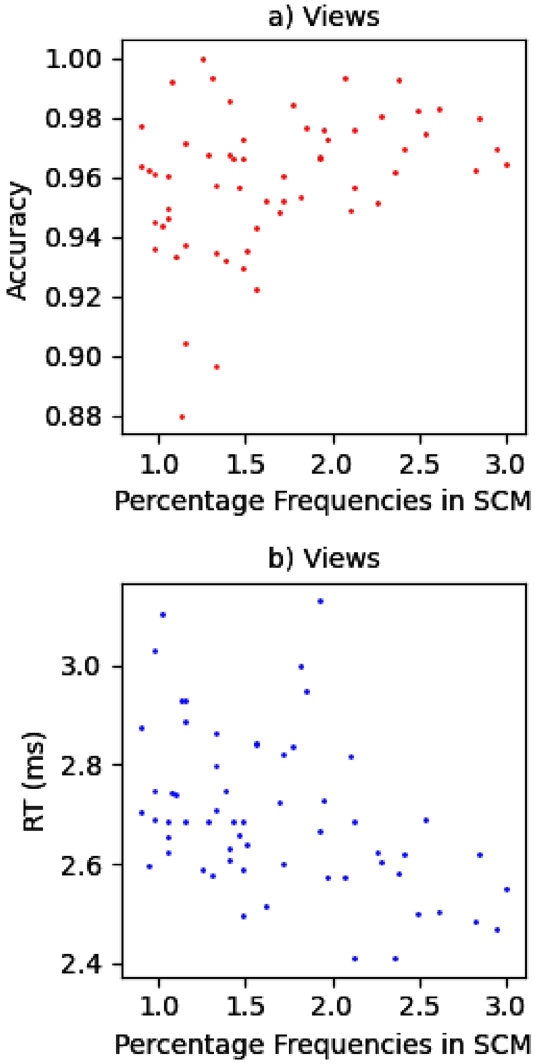


Figure 3.9. (a) The relation between the action recognition accuracies of the starting views and the percentage frequencies of the ending views. (b) The relationship between the action recognition RTs of the starting views and the percentage frequencies of the ending views.

3.4 Conclusions

The first major goal of this pilot study was to unveil whether the action recognition performance of human observers with MVVHAs under active viewpoint movements is higher than under passive viewpoint movements. Thus, I compared the action recognition performance of the participants in SCM trials with active viewpoint movements and in RM trials with passive random viewpoint movements. The design of this comparison was within-subject. My first hypothesis was that if human observers select the viewpoint efficiently for action recognition, then their accuracy would have been higher or their RT lower when they can select the views like in SCM condition than when they cannot select the views like in RM condition. Action recognition in the SCM condition was not significantly more accurate than in the RM condition. This may be due to the fact that it was very high in all conditions and it was not possible to find any significant difference in accuracy between conditions of any independent variables. Therefore, the conclusions in this section about action recognition performance were not based on differences in accuracy between conditions. They were only based on differences in RT between conditions. The RT of action recognition in SCM was significantly faster than in RM. This speed improvement might have been due to the fact that human observers are able to select views efficiently for action recognition and they could only use this advantaging skill only in SCM and not in RM.

The second main goal was to investigate whether humans select the efficient views more often than the inefficient viewpoints during active action recognition. This led me to two further subordinate objectives. The first one was to determine whether the action recognition performance with MVVHAs is viewpoint-dependent. The distinction between efficient and inefficient viewpoints is only valid if the recognition performance is viewpoint-dependent and thus the recognition performance significantly varies across different viewpoints. The results showed that the action recognition accuracy was influenced by the elevation angle ϕ of the starting viewpoints and not by the azimuth angle θ. Additionally, the action recognition RT was affected by both the elevation angle ϕ and the azimuth angle θ. Therefore, the action recognition performance was viewpoint-dependent, and this supports the validity of efficient and inefficient viewpoints. The action recognition performance might have been viewpoint-dependent because the recognising action categories were similar to each other. The literature review in chapter 4 includes many studies that found a similar pattern for object recognition. These studies (Hayward & Williams, 2000; Leek & Johnston, 2006; Tarr & Hayward, 2017; Tarr & Pinker, 1990; Tjan, 2001) highlighted that the object recognition of their human participants was viewpoint-dependent when they recognised similar objects and it was viewpoint-invariant when they recognised very different objects.

The second subordinate objective was a follow-up of the first subordinate objective. If the action recognition performance with MVVHAs is viewpoint-dependent, then the second subordinate objective was to highlight the efficient and the inefficient viewpoints for the action recognition of humans. I assumed that the performance of action recognition is better if participants start watching the action from the efficient views than from the inefficient views. Therefore, I looked at the efficient angle ϕ and efficient angles θ. Speaking of efficient angles ϕ, the middle-top views with $\phi=60°$ are the most efficient because action recognition was the most accurate and the fastest when participants started watching the actions from middle-top views. On the other hand, the bottom views with $\phi=150°$ are inefficient because action recognition was less accurate and longer with bottom starting views. Turning the argument into the efficient angles θ, the efficient angles θ were the side (right and left) angles θ and side-front (right-front and left-front) angles θ as classification was more accurate and faster with these starting angles θ. The back, side-back and front angles θ were inefficient given that the action recognition was less accurate and longer when the participants started watching the actions from these view angles θ.

By going back to the second main goal, the human participants selected the efficient views more often than the inefficient views for their action recognition because the accuracies of the views were positively correlated with the selection of the views, while the RTs of the views were negatively correlated with the selection of the views. In fact, if some viewpoints are efficient, then participants score higher accuracy and shorter RT from these viewpoints. Therefore, if participants select more often the efficient views, the selection of these efficient views should be higher. However, if some views are inefficient, then the accuracies from these views are lower and RT longer. In this case, the selection of these inefficient views would be lower, if humans select the efficient views. Thus, I expected the selection of views positively correlated with the accuracies of the views and negatively correlated with RTs of the views. The results of this study revealed this pattern, and I can conclude that they selected the efficient views more often than the inefficient views.

Given that participants improved their performance in action recognition when they could move their own view in the SCM condition and given that they selected the efficient views more often than the inefficient views, then the active action recognition of human observers is efficient.

4 Efficient and Inefficient Views for the Action Recognition of Human Observers

4.1 Introduction

The main objective of the pilot study of the previous chapter 3 was to discover whether the action recognition performance (accuracy and RT) of humans is better in the SCM condition than in the baseline RM condition. Even if the accuracy in the SCM trials was not significantly higher than in RM trials and the RT in the SCM trials was significantly faster than in the RM trials. Based on these results, I concluded the performance of the human participants in the SCM was a marginal better than in RM. By analysing the same data of the pilot study, I was also able to show that participants select the efficient views more often than the inefficient views. However, there was a substantial limit in the way I discriminate the efficient and inefficient views because their discrimination was not the main objective of the study, and the study was not specifically designed for it. I used the accuracy and the RTs of the starting viewpoints to assess whether they were efficient or inefficient. Yet, the accuracy and the RT in each trial did not rigorously depend on a single view, given that the view changed within each trial in both SCM and RM conditions. Therefore, the results may have been corrupted. This limit may have led me to wrong conclusions. To tackle this issue, I made the study of this chapter with no view movements within the trials to strictly reveal the efficient and inefficient viewpoints.

The distinction of efficient and inefficient viewpoints for action recognition is valid only if the action recognition performance is viewpoint-dependent, meaning that the performance is affected by the observer's viewpoint at the time of action observation. The study of Mitchell and Curry (2016) nearly addressed this issue despite of some limitations. They did not specifically study whether the action recognition is viewpoint-dependent, but they examined whether the actor recognition is viewpoint-dependent. They asked participants to recognise some actors by the point-light displays of their walks from two different viewpoints. The two views were the front (θ = −45°) and right-front views (θ = −45°). There was no significant difference in accuracy in the front and in the right-front conditions. However, this study cannot clearly address my aim of this chapter because it had three issues. One, they studied actor recognition rather than action recognition. Two, they only used two views while there are potentially infinite viewpoints in the real world. Three, they did not report the RTs of the participants. Therefore, in the study of this chapter, I will tackle these issues, by directly looking at the action recognition, by using fifteen different views of my own MVVHA, and by recording both the action classification accuracies and RTs of the human participants.

Because there is lack of substantial studies and theories on whether the action recognition is viewpoint-dependent we could infer it from the literature addressing whether object recognition is viewpoint-dependent or viewpoint-invariant. There has been a long debate (Biederman & Bar, 1999, 2000; Biederman & Gerhardstein, 1993, 1995; Hayward & Tarr, 1997, 2000; Tarr & Bulthoff, 1993, 1995, 1998) between two broad classes of theories explaining how the brain recognizes objects across the infinite possible chances in viewpoint. One class includes the 2d theories and the other one consists of the 3d theories. Both types of theories agree that object recognition is the process of matching the encoded representations of the perceived objects with the encoded representations of objects stored in the memory. However, they have mostly disagreed on whether the encoded object representations are viewpoint-dependent or viewpoint-independent.

The view-based theories typically argue that an object is encoded by a set of 2d image-based and view-based representations, one for each pose or viewpoint (Bricolo et al., 1997; Logothetis et al., 1994; Poggio & Edelman, 1990; Rock & Divita, 1987; Tarr & Bulthoff, 1995, 1998; Tarr & Pinker, 1989). For instance, Lawson and Humphreys (1998) concluded that “object recognition is mediated by stored representations that are both view- and object-specific.” These 2d representations are viewer-centred. Thus, the 2d theories claim that 2d object representations are viewpoint-dependent and hypothesise that there should be a significant variation in the object recognition performance across viewpoints. Specifically, they predict that the recognition performance from familiar viewpoints is higher than from unfamiliar viewpoints because the 2d image-based and view-based information which are essential for higher recognition performance has been only learnt and memorized from the familiar viewpoint and never from the unfamiliar ones.

In contrast, according to the 3d theories, an object is represented by a specific combination of 3d representations which are object-centred and thus are viewpoint-independent or viewpoint-invariant (Biederman, 1987; Biederman & Gerhardstein, 1993; Hummel & Biederman, 1992; Marr, 1982; Marr & Nishihara, 1978). Therefore, the 3d theories predict that object recognition from unfamiliar viewpoints is as efficient as from familiar viewpoints because 3d structural descriptions do not depend on the viewpoint by definition.

However, the empirical results have also been controversial. These researchers have assumed that large variations in object recognition performance, in terms of RT or accuracy, across different viewpoints are evidence that object recognition depends on 2d viewpoint-dependent representations. On the other hand, they also presumed that insignificant variations in the object recognition performance across different viewpoints are proof that object recognition involves 3d viewpoint-invariant representations. However, the results are controversial because some experimental results showed viewpoint-dependent object recognition performance (Tarr & Bülthoff, 1999), whereas others indicated viewpoint-independent object recognition performance (Bart & Hegdé, 2012). Therefore, object recognition performance appears to be viewpoint-dependent or viewpoint-invariant, depending on the circumstances such as stimuli and tasks and more.

As far as I am concerned, showing that object recognition performance of humans is viewpoint-dependent or viewpoint-independent does not necessarily prove that object representations of humans are 2d, 3d or even both. For example, the representations involved in object recognition are not inevitably 2d even if all recognition performances were viewpoint-dependent. Theoretically, the object recognition may still depend on 3d representations. The 3d theories postulate that that the 3d viewpoint-invariant representations are produced by several processes transforming the 2d information that fall on the retina into 3d encoded object representations. Therefore, it is possible to argue that the processes producing the representations (not necessarily the representations) may depend on the viewpoint and these processes alone could cause the view-dependent recognition performance even if the produced representations may still be in 3d and may not depend on the viewpoint. The object recognition could be more accurate and faster from familiar viewpoints than from unfamiliar ones just because the processes producing the representations familiar viewpoints may completely or partially different than from unfamiliar viewpoints. A plausible difference of the processes may be their degree of automaticity (Moors, 2016; Moors & De Houwer, 2006). In fact, because of more practice from the familiar viewpoints, the processes forming the representations from familiar viewpoints might become automatic and so fast and more accurate, while the ones from unfamiliar viewpoints might still be controlled and so slow and less accurate. Thus, conclusion that that the representations are 2d is not inferable from the evidence of viewpoint-dependent object recognition performance.

Likewise, we cannot conclude that the representations are undoubtedly 3d even if all recognition performances were viewpoint invariant. They may still be 2d. One potential reason could be that when learning recognising a new object from specific viewpoints, the brain may also estimate the 2d image-based representations of the object from the other never-used viewpoints by some transformations. Therefore, even if the brain has only seen that object from the familiar viewpoints, the unfamiliar viewpoints might also become familiar because the image-based representations of the object from the unfamiliar viewpoints were generated by some transformations, stored in the memory and ready to be used to recognise that object from the unfamiliar viewpoints as accurately and quickly as from the familiar viewpoints. In this way, the object recognition performance could be viewpoint-invariant even if the involved representations are only in 2d.

To tackle the controversial results, some recent models explain how different components of object recognition could prompt viewpoint-dependent and viewpoint-invariant object recognition behaviours under different circumstances. These various models encompass a wide range of perspectives. Hummel and Stankiewicz (1996), and others (Hummel, 2001; Stankiewicz et al., 1998), affirmed that object recognition primarily relies on two structural representations: one is viewpoint-invariant and the other is viewpoint-dependent. The view-invariant representation, the independent geon array (IGA), is a collection of units represent the object’s parts (geons). Each geon is represented by one unit independently of every other. The IGA are viewpoint-invariant because of this independence. The viewpoint-dependent representation, the substructure matrix (SSM), is also a collection of units that represents the geons at each location of a coordinate system which is semiobject-centred. Therefore, the geons in the SSM are not represented independently but dependently on their relative locations. Consequentially, SSM depends on the viewpoint and mirror reflections. They also claimed that the IGA requires visual attention and processing time while the SSM can be processed without these processing resources. Then, object recognition in lack of processing resources is viewpoint-dependent because it mediated by the SSM, the viewpoint-dependent representation that do not require attention and time, while the object recognition with available resources is viewpoint-independent because of IGA. Other alternative theories have been proposed. Edelman and Intrator (2003) suggested an 2d image-based model that can capture certain elements of 3d object structure. Cooper and Wojan (2000) proposed the utilization of either viewpoint-invariant or viewpoint-dependent information for different tasks. Tjan (2001), Foster and Gilson (2002), and later Tarr and Hayward (2017) proposed the independent and concurrent encodings of both viewpoint-dependent and viewpoint-independent information.

I support the independent encoding of multivariate viewpoint-dependent and viewpoint-independent object features as proposed by Tjan (2001), Foster and Gilson (2002), and later Tarr and Hayward (2017). Through evolution, humans, as well as many another animal, have developed many types of sensory systems to perceive the surrounding objects and assess them as hostile or harmless. The sensory systems are the visual, auditory, tactile, olfactory, gustatory systems and even more. Diversifying the perception resources with multiple sensory systems has beneficial for humans and many other animals. One advantage of diversifying the perception resources with multiple sensory system is to preserve the ability to effectively interact with the world even if one sensory system gets damaged. it is more likely that only one sensory system is completely damaged in the one person than multiple systems at the same time. By having multiple sensory systems, humans with any damaged sensory system can still perceive objects to some degree from the other intact sensory systems. If they had only the visual system and this gets damaged, they could not perceive anything from the world and could totally lose their ability to effectively interact with it. Another advantage of the sensory system diversifications is that is to perceive the objects accurately and fast even if one type of signals is disrupted by some environmental circumstances. It is more likely that the environmental circumstances (dark) disrupt one single type of signals (visual) than multiple circumstances (dark and loud noise) disrupt multiple signal types (visual and auditory) at the same time. Thus, humans with multiple sensory systems can still perceive the objects from only the auditory signals in dark and quiet environments and from visual signals in bright and loud environments. Instead, if they only had the visual system, their object perception could be totally impaired by the dark. My claim here is that as humans developed multiple sensory systems to diversify the perception resources, they may also have developed different subsystems within the same visual system that encodes both the viewpoint-dependent and viewpoint-invariant visual features of the objects.

Researchers have also shown the circumstances which promote viewpoint-dependent or viewpoint-independent object recognition performances. Stankiewicz, Hummel, and Cooper (Stankiewicz et al., 1998) showed the involvement of attention. They revealed that the increase of object recognition performance for priming was reflection-dependent for unattended objects and reflection-invariant for attended objects. By assuming the similarity between reflection and rotation, the study indicates that object recognition performance is viewpoint-invariant if the objects are attended. Otherwise, the performance is viewpoint-dependent if the objects are unattended.

Milivojevic (2012) noted that the handed (mirrored) object recognition, such as the recognition of the left shoes and the right shoes, is highly view-dependent, while non-handed (non-mirrored) object recognition, such as the recognition of shoes and bottles, is viewpoint invariant. They also claimed that the perception of the spatial relations between objects' features is affected by inversion while the perception of the object features themselves is not disrupted by inversion. Therefore, the mirrored object recognition is viewpoint-dependent because the mirrored objects have the same features and can only be differentiated by the spatial relations of their features. However, the perception of these spatial relations is impaired by inversion. On the other hand, the non-handed object recognition is viewpoint-independent because the non-handed objects are distinguishable by both their independent features and their spatial relations, and the perception of the independent features is not affected by the inversion.

Others (Biederman & Gerhardstein, 1993; Hamm & McMullen, 1998; Milivojevic, 2012; Tarr & Bulthoff, 1995) argued that the level of the recognised object classes determines the viewpoint dependency of the recognition performance. Low-level object recognition is the recognition of the subordinate and specific object classes like dog breeds, while high-level recognition is the discrimination of superior and general object classes like animals and plants. The performance of low-level object recognition is viewpoint-dependent, whereas the performance of the high-level recognition is viewpoint-invariant. Rosch and colleagues (1976) described the low-level and high-level object recognitions by being at subordinate-level and at basic or entry level, respectively. In fact, Yin (1969) highlighted the recognition of faces as belonging to a specific person is highly disrupted when the faces are inverted. According to this approach, this is because the recognition level is very subordinate for faces and therefore viewpoint-dependent. Additionally, Corballis and colleagues (1978) found that the identification of letters of the alphabet (subordinate-level recognition) was disrupted by character orientation, while the performance was viewpoint-invariant for the classification of alphanumeric characters as being letters or numbers (superordinate-level recognition). Moreover, Hamm and McMullen (1998) found that subordinate-level recognition such as recognition of dog breeds is impaired by changes in viewpoint while basic level recognition like dog recognition is not affected by viewpoint changes.

Hayward and Williams (2000), and Tjan (2001) agreed that the viewpoint dependency and the recognition level are highly correlated. However, they stressed that the shape similarity (not the category level) of the recognised objects is the real cause of the viewpoint-dependent object recognition performance. They also noted that the reason why object recognition performance tends to be viewpoint-dependent with subordinate object categories is that that the shape similarity is generally high for subordinate object categories.

However, the framework of Tarr and Pinker (1990) and some others (Leek & Johnston, 2006; Tarr & Hayward, 2017) more precisely defines the similarity of objects. If the orders of objects' parts are distinguishable along the top-to-bottom axis, then the objects are not similar and can be easily recognised independently from the viewpoint. However, when the part orders of the objects, like mirrored objects and others, are not different along the top-to-bottom axis and the second left-to-right axis is required to be differentiated, then these objects are similar, and the recognition performance is viewpoint dependent.

If the same principles of object recognition applied to the action recognition, then the action recognition performance would be viewpoint-dependent for the recognition mirrored actions (like waving with left hand vs. waving with right hand) and the recognition of similar actions like discussing with hand gestures, waving and pointing. However, although the study of chapter of this thesis did include these types of actions, I did not investigate the conditions when the action recognition performance is viewpoint-dependent and the conditions when it is viewpoint invariant, because this was not the aim of this chapter.

The study of this chapter had two main aims. The first one was to demonstrate that the action recognition performance of people was viewpoint-dependent when their task is to passively recognise the actions classes of MVVHAs. We can only find evidence of efficient active action recognition when action recognition is viewpoint-dependent. In this way, we could test whether people select the efficient viewpoints more often than the other inefficient ones. It would not be possible when the action recognition is viewpoint-invariant because each viewpoint is as efficient as all other ones and there is no distinction of efficient and inefficient viewpoints. The second aim was a follow-up of the first one. This was to unveil efficient and inefficient views for humans when they passively recognize the MVVHAs. The major reason is that knowing which viewpoints are efficient and which ones are inefficient, then I could test whether humans select their efficient views more often than their inefficient ones.

To evaluate that the recognition of MVVHAs is viewpoint-dependent and then reveal the efficient views and inefficient views, I asked human participants to passively classify the same actions from different views and look at whether the action recognition accuracies and RTs of all views are significantly different. I will conclude that the action recognition with MVVHAs is viewpoint-dependent if action recognition accuracy or RTs are significantly affected by the viewpoint. I will conclude that some viewpoints are efficient for humans if the human participants classify the actions more accurately or faster from these views rather than from the others. I will also conclude that some other viewpoints are inefficient for humans if the human participants classify the actions less accurately and more slowly from these views than from the other views.

A hypothesis of the study was that the action recognition performance with MVVHA would have been viewpoint-dependent and therefore the accuracy or RT would have significantly varied across different viewpoints. This was because the actions classes in MVVHA are similar to each other. Discussing, pointing and waving have similar shape as well as falling and dancing. Sitting down and standing up are basically the same if we do not consider inversion in time. A confusion matrix was also planned to show how similar the actions classes are. The actions are more similar should be confused more often. I did not make any predictions about which views are efficient and which ones are inefficient. My objective was only to identify them such that in the future studies, I can verify whether the human observers select more often the efficient views rather than the inefficient ones in the active conditions.

4.2 Methods

4.2.1 Participants

I recruited 49 participants who were undergraduate students at the University of Essex. They participated online via a web browser. In order to ensure the participants were following the instructions and paying attention throughout the task, I used several criteria to check performance. I excluded 3 participants because one of following conditions or more was true:

- Their action classification accuracy was below 0.4;
- Any of their local accuracy was less than 0.2. The local accuracy is accuracy scored in any 20 neighbouring (or consecutive) trials;
- More than 6 trials out of 120 (5%) were excluded. I excluded all trials with RT either lower than 250 ms or larger than 7,000 ms;
- More than 3 trials out of 20 (15%) were excluded in any 20 neighbouring trials.

Therefore, I only analysed 46 participants. From now on, any statements about the participants of this study will be referred to the 46 non-excluded participants unless specified differently. Their age had a mean of 20.41 years and ranged from 18 to 41. 12 participants were male and the other 34 were female. I rewarded the participants with academic credits.

4.2.2 Materials

For this online study, I used the smaller version of my dataset MVVHA which I named o.3.6 (see full description in chapter 2). The version ob.3.6 contains 1512 MVVs. Each MVV includes 24 SVVs showing the same actor performing the same action from 24 different viewpoints. Thus, in total, there are 36,288 SVVs (1512 MVVs x 24 views). The images were blurred and stored in a GitHub repository (https://github.com/ccalafiore/dataset_ob.3.6). Transparency was added on-the-fly when images were shown to the participants. All images were shown blurred and transparent to make the action recognition harder for the participants.

The dataset version ob.3.6 has all 7 classes of actions: dancing, discussing, sitting down, standing up, falling down, pointing, waving. There are 27 different actions in each action class. Furthermore, each animation had the mirrors and the non-mirrored conditions. There are 4 different actors. Every SVV of ob.3.6 has 10 frames which were refreshed at 5 Hz making each video 2 seconds long. Each frame is in colour and has size of 224 x 224 x 3 pixels.

There are only 8 angles θ and 3 angles ϕ which made 24 views. However, as a result of pilot experiments and constraints over online presentation, I did not use all 36,288 SVVs for this experiment. I only used 22,680 SVVs (1512 MVVs x 15 views), by only choosing 15 views out of 24. In fact, I only used the SVVs with specific 5 angles θ of ob.3.6: −180° (back), −135° (right-back), −90° (right), −45° (right-front), 0° (front). I did not include the left side views because I showed the videos of both mirror conditions and so I assumed that participants would perform approximately the same from the left side views and right side views. However, I included all 3 angles ϕ: 45° (top), 90° (middle), 135° (bottom). Therefore, I selected 15 views (5 angles θ x 3 angles ϕ) for this experiment.

The online test was made as a Qualtrics survey (www.qualtrics.com). Only one question of the survey runs the action recognition experiment. This experiment question had an html frame that displays the stimulating website that I built with GitHub Pages (pages.github.com). This website runs the stimulus presentation script which is an html file with html and JavaScript code. You can visit this website at https://ccalafiore.github.io/demos/demo_1 for an example of the stimuli and the task. The experiment was controlled using the library jsPsych (www.jspsych.org) which is designed for running experiments in a web browser and has good timing performance.

4.2.3 Design

I used a 5 x 3 within-subject design. The two independent variables were the angle θ and angle ϕ of the views from which the participants look at the action. I only used 5 angles θ of ob.3.6: −180° (back), −135° (right-back), −90° (right), −45° (right-front), 0° (front). However, I used all 3 angles ϕ of ob.3.6: 45° (top), 90° (middle), 135° (bottom). The dependent variables were the action recognition accuracy and the action recognition RT taken from the start of the animation.

4.2.4 Procedure

Each participant went through 1-section online study of about 20 minutes. I provided to each participant a web link to run the experiment online. They were instructed to open the link with Google Chrome or Microsoft Edge on their own laptop or desktop PC. First, they agreed to the consent form. Then, they responded to 2 questions asking them their age and gender. Finally, they started the action recognition task which had 2 phases: familiarization and real test. For each participant, I randomly selected 4 classes of action out of 7, 2 actors out 4 and 2 animations per class out of 27. This means I selected 8 animations (4 classes x 2 animations per class) at random. I showed SVVs that had the same 4 action classes in both the familiarization and the real test. However, I showed different actors and different animations per class in the familiarization and real test. Participants could only see the videos of one of the 2 randomly selected actors in the familiarization and the other one in the real test. Likewise, they could only see one of the two random animations per class in the familiarization and the other one in the real test. This was designed so that participants could familiarise themselves with exactly the style of animation that they would see but without being exposed to the exact same exemplars.

In both the familiarization and the real test, each trial started with a white fixation cross in the centre of the screen for 1 second plus a jitter of 0.5 seconds. Then, I showed the 10-frame video at 5Hz for a maximum of 2 seconds. The blurred images were made transparent on-the-fly. If they did not classify the action within the 2 seconds, I showed a screen with the text “Which Action?” until their classification. They could classify the action by pressing one of the following keys: “1”, “2”, “3”, “4”. I showed the keys next to its corresponding action class below the video during the whole experiment to remind the participants. I showed each pair of key and class with a different colour. I randomised both the order and the colour of the classes on the screen. However, the order and the colour of the classes on the screen were kept the same for each trial in both the familiarization and real test to avoid confusing the participants. Only in the familiarization, just after their classification, I displayed for 1 second a feedback screen saying “Correct!” in green if the classification was correct or “Incorrect!” in red if it was wrong. In the real test, this feedback was not shown, so this feedback screen was skipped. After the feedback or response, the next trial started with the fixation cross.

The familiarization only contained 8 action classification trials. The 4 classes of actions were shown 2 times from a random view and with a random mirror condition. I did not analyse the data from the familiarization. The aim of the familiarization was just to show participants how the test looked and to give some practice with the task. I only analysed the real test phase which could be spilt into 4 sequential blocks of 30 trials. Thus, there were 120 trials in total in the real test (4 classes x 1 actors x 1 animation per class x 2 mirror x 5 angles θ x 3 angles ϕ). The order of the trials was randomised in both the familiarization and real test.

The instructions below were shown to the participants before starting both familiarization and the real test.

> *"You will go through two phases: Familiarization and Real Experiment. In both phases, you will do several trials. In each trial, your main task is classifying the action displayed in an unclear video of 2 seconds. You can respond anytime during or after the video plays.*
>
> *Each trial starts with a white fixation cross at the centre of the screen for 1 second. Here, you are asked to look at the cross because that is the position where the action is about to happen.*
>
> *Next, the screen will display a video of a person doing 1 of the 4 possible actions until your action classification or for 2 seconds. Here, you are asked to classify the action by pressing 1 for Falling, 2 for Waving, 3 for Sitting Down, 4 for Pointing. In case you forget the names of the actions, they will always be shown on the bottom of the screen with their numbers*

during the whole experiment. Please, classify as quickly as possible.

If you do not respond during the 2 seconds, you will see a new screen saying Which Action? until you classify. Here, you are again asked to classify the action as quickly as possible. If you did not recognise any action in the unclear video, please have a guess as quickly as possible.

In the Familiarization, just after your classification, you will get a feedback for 1 seconds saying Correct if your classification was correct or Incorrect if your classification was incorrect. In the Real Experiment, there is no feedback and this step is skipped.

Next, a new trial will start with the white fixation cross."

You can run the action recognition test at https://ccalafiore.github.io/demos/demo_1.

4.3 Results

Overall, the action classification accuracy of the human participants was 0.76 (*SD* = 0.125) and their RT was 2296 ms (*SD* = 389 ms). Note that the chance of correct classification was not .143 (1 out of 7), even if the total action classes across all participants were 7. The chance was actually .25 (1 out of 4) since the test randomly selected only 4 out 7 action classes for each participant.

4.3.1 Learning Effect

The charts in the Figure 4.1 summarise the learning effect throughout the whole test, by showing the action recognition accuracies and the RTs in each block of 30 trials. Both accuracies and RTs show signs that steady leaning at each block, though the sings are more evident in the RTs than in the accuracies. The accuracy steadily increases whereas the RT progressively decreases by increasing the number of the blocks of trials.

I ran a one-way ANOVA to look at the effect of the blocks of trials on the action recognition accuracy and RT. All effects are reported as significant at *p* < .05. Mauchly's test indicated that the assumption of sphericity had been violated for the effect of blocks on RT, $\chi^2(5)$ = 36.853, *p* < .001. Therefore, the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (*ε* = .651).

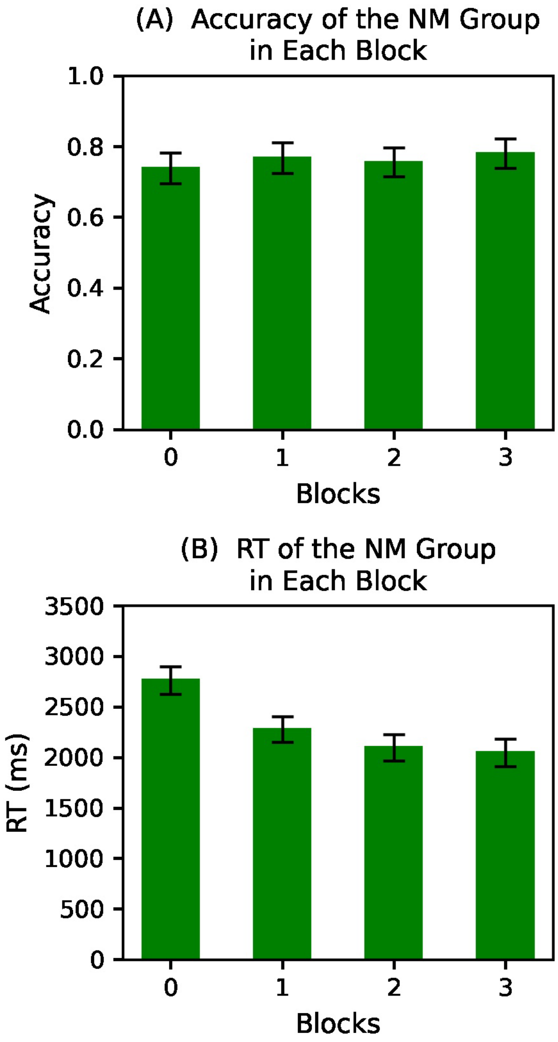


Figure 4.1. Action classification accuracies (A) and RTs (B) of the human participants in each block of 30 trials.

There was no significant effect of the blocks on accuracy, *F*(3, 135) = 2.270, *p* = .083. Because the p-value was nearly .05, I followed up on this effect anyway. I ran 6 pairwise paired t-tests with Bonferroni correction to compare the accuracies of the participants in all 4 blocks of trials. Even though no t-test showed a significant difference between the blocks because of the conservative correction, action recognition in the first block was the least accurate (*M* = .738, *SD* = .144) and increased in the second block (*M* = .768, *SD* = 0.147). Next, the accuracy was slightly decreased in the third block (*M* = .755, *SD* = .139) and it raised in the last block (*M* = .780, *SD* = .142) where it was the highest.

The was a significant effect of blocks on RT, *F*(1.954, 87.919) = 74.017, *p* < .001. I followed up on these effects, by running 6 pairwise paired t-tests with Bonferroni correction to compare the RTs of the participants in all 4 blocks of trials. There were significant differences of RTs in five out of six pairwise comparisons. Only the difference in RTs of the third (*M* = 2,097 ms, *SD* = 442 ms) and fourth (*M* = 2,047 ms, *SD* = 464 ms) blocks was not significant, *t*(45) = 1.361, *p* = 1.0.

4.3.2 Efficient Views and Inefficient Views

In this study, I aimed to highlight which views are efficient and which ones are inefficient for action recognition. The action recognition performance should be significantly higher from the efficient views than from the inefficient views. In other words, the action recognition should be more accurate and faster from the efficient views than the inefficient views. Thus, I manipulated the view to look at whether it affects the accuracy and the RT of action recognition.

Firstly, I run a one-way ANOVA to unveil the effect of the view on the action recognition accuracy and RT of the human

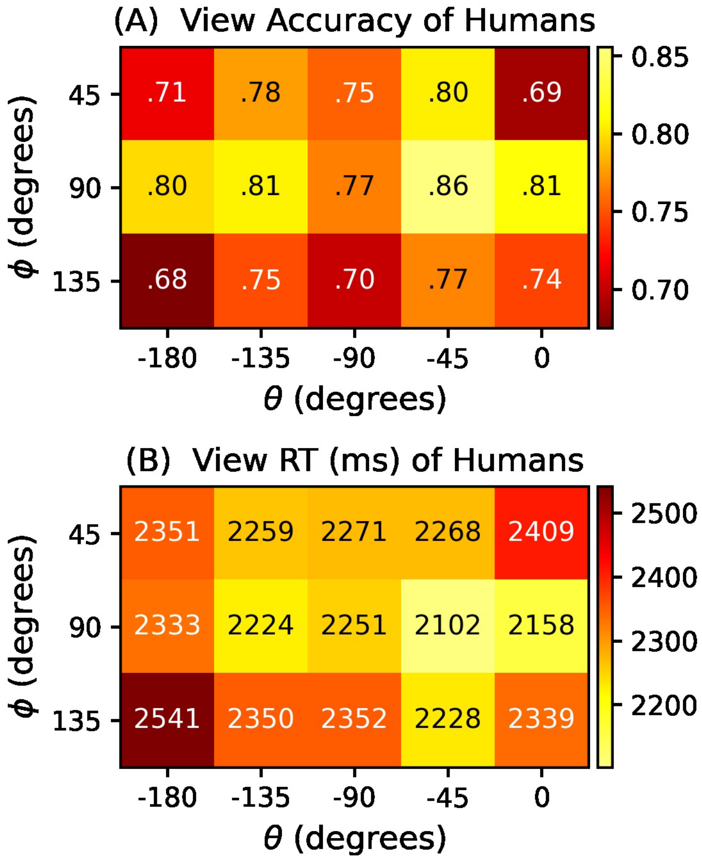


Figure 4.2. Action classification accuracies (A) and RTs (B) of the human participants in all views. These views were definened by the two angles θ and ϕ.

participants. There were 15 within-participant conditions (5 angles θ x 3 angles ϕ = 15 views). All effects are reported as significant at $p < .05$. Mauchly's test indicated that the assumption of sphericity had been violated for the effect of views on both accuracy, $\chi^2(104) = 143.103$, $p = .008$, and RT, $\chi^2(104) = 174.799$, $p < .001$. Hence, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .671$ for the effect of the views on accuracy and $\varepsilon = .643$ for the effect on RT).

The ANOVA showed a significant effect of view on accuracy, $F(9.390, 422.538) = 6.045$, $p < .001$, and a significant effect on RT, $F(9.002, 405.094) = 7.171$, $p < .001$. I did not follow up on the effect of views by pairwise t-tests because of the numerous (105) possible pairs of views. Anyway, the heatmaps in Figure 4.2 illustrate the action classification accuracies (A) and the RTs (B) of the participants from every viewpoint. What stands out from the heatmaps is that action recognition is more accurate and faster from the middle (ϕ=90) views than from the top (ϕ=45) and bottom (ϕ=135) views. In addition, the right-front (θ=−45) views also tend to be more accurate and faster than the others.

I conducted a 5 x 3 two-way ANOVA to reveal the main effects of angle θ and of angle ϕ, and the interaction effect of them on accuracy and RT. All effects are reported as significant at $p < .05$. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of angle θ on accuracy, $\chi^2(9) = 44.507$, $p < .001$, and for the interaction effect of angle θ and angle ϕ on RT, $\chi^2(39) = 51.620$, $p = .036$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .660$ for the main effect of angle θ on accuracy and $\varepsilon = .788$ for the interaction effect of angle θ and angle ϕ on RT).

There was a significant main effect of angle θ on both accuracy, $F(2.642, 118.886) = 6.440$, $p < .001$, and RT, $F(4, 180) = 8.465$, $p < .001$. I followed up on the effects by running multiple paired t-tests with Bonferroni correction to compare the five conditions of angle θ in both accuracy and RT. Action recognition was significantly less accurate from the back views θ$_0$=−180° (*M* = .729, *SD* = .149) than from the right-back views θ$_1$=−135° (*M* = .777, *SD* = .148), *t*(45) = −3.053, *p* = .038, and from the right-front views θ$_3$=−45° (*M* = .809, *SD* = .151), *t*(45) = −4.699, *p* < .001. Likewise, action recognition from the back views θ$_0$=−180° (*M* = 2,409 ms, *SD* = 413 ms) was significantly slower than from the right-back views θ$_1$=−135° (*M* = 2,278 ms, *SD* = 466 ms), *t*(45) = 3.275, *p* = .021, the right views θ$_2$=−90° (*M* = 2,291 ms, *SD* = 401 ms), *t*(45) = 3.046, *p* = .041, from the right-front views θ$_3$=−45° (*M* = 2,200 ms, *SD* = 411 ms), *t*(45) = 6.406, *p* < .001, and from the front views θ$_4$=0° (*M* = 2,301 ms, *SD* = 402 ms), *t*(45) = 3.528, *p* = .011. On the other hand, action recognition from the right-front views θ$_3$=−45° (*M* = .809, *SD* = .151) was significantly more accurate than from the back views θ$_0$=−180° (*M* = .729, *SD* = .149), *t*(45) = 4.699, *p* < .001, from the right-back views θ$_1$=−135° (*M* = .777, *SD* = .148), *t*(45) = 3.092, *p* = .036, from the right views θ$_2$=−90° (*M* = .740, *SD* = .160), *t*(45) = 3.505, *p* = .011, and from the front views θ$_4$=0° (*M* = .748, *SD* = .124), *t*(45) = 3.686, *p* = .006. See Figure 4.3(A) for the accuracies of the

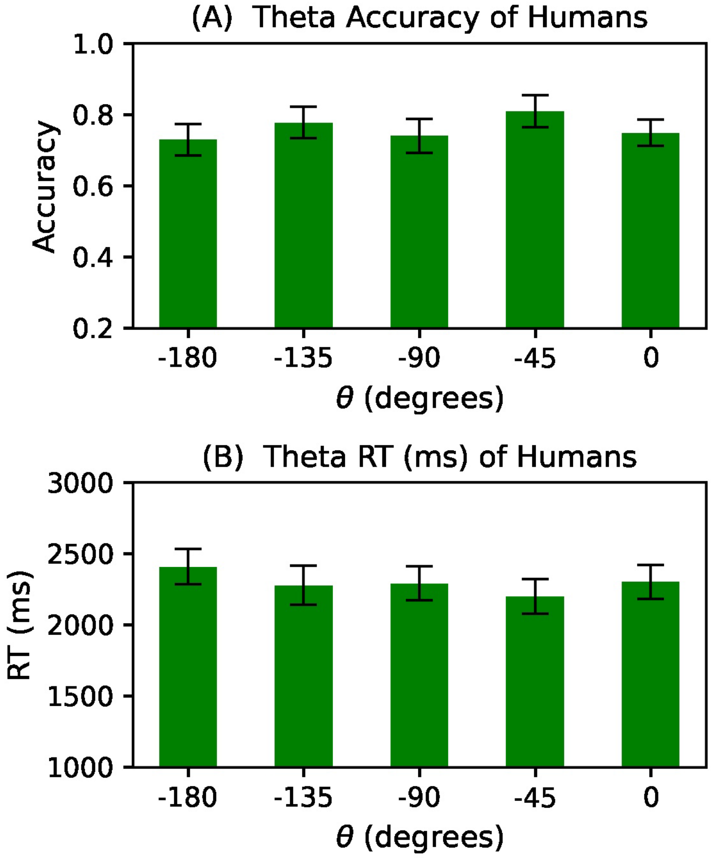


Figure 4.3. Action classification accuracies (A) and RTs (B) of the human participants in each viewpoint angle θ.

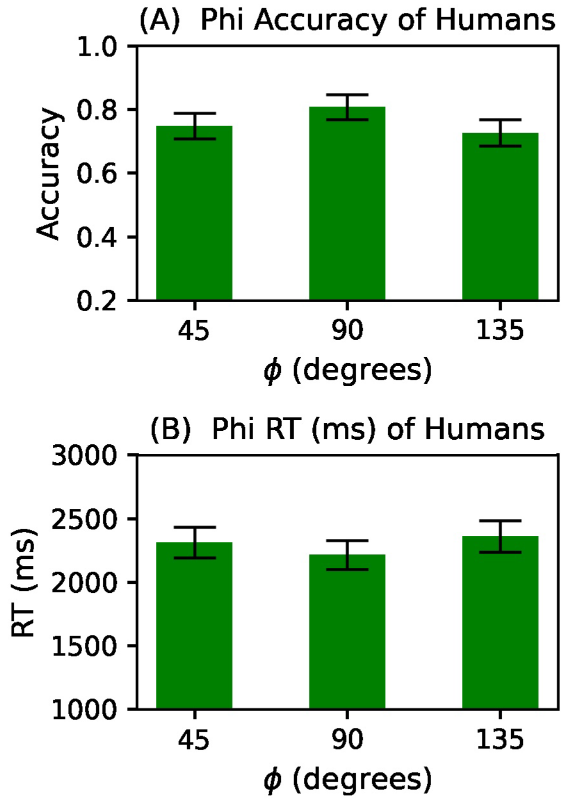


Figure 4.4. Action classification accuracies (A) and RTs (B) of the human participants in each viewpoint angle ϕ.

participants in all five viewpoint angles θ. Furthermore, RT was significantly faster from the right-front views θ_3=−45° (*M* = 2,200 ms, *SD* = 411 ms) than from the back views θ_0=−180° (*M* = 2,409 ms, *SD* = 413 ms), *t*(45) = −6.406, *p* < .001. The short action recognition RT in the right-front views θ_3=−45° (*M* = 2,200 ms, *SD* = 411 ms) was nearly significant with respect to the right-back views θ_1=−135° (*M* = 2,278 ms, *SD* = 466 ms), *t*(45) = −2.716, *p* = .089, from the right views θ_2=−90° (*M* = 2,291 ms, *SD* = 401 ms), *t*(45) = −2.615, *p* = .115, and from the front views θ_4=0° (*M* = 2,301 ms, *SD* = 402 ms), *t*(45) = −2.924, *p* = .051. See Figure 4.3(B) for the RTs of the participants in all five viewpoint angles θ.

To summarise the results as a function of angle θ, performance was poorer (less accurate and slower) on the back views. Interestingly, the front views (θ_4=0°) did not show the best performance, but rather a viewpoint slightly offset to the right (θ_3=−45°).

There was also a significant main effect of angle ϕ on accuracy, *F*(2, 90) = 18.014, *p* < 0.001, and RT, *F*(2, 90) = 15.682, *p* < 0.001. I further investigated these main effects with multiple paired t-tests with Bonferroni correction between three conditions of angle ϕ. Action recognition from the middle views ϕ_1=90° (*M* = .808, *SD* = .132) was significantly more accurate than from the top view ϕ_0=45° (*M* = .747, *SD* = .137), *t*(45) = 4.105, *p* <

.001, and from the bottom views ϕ_2=135° (*M* = .727, *SD* = .140), *t*(45) = 5.855, *p* < .001. Additionally, action recognition from the middle views ϕ_1=90° (*M* = 2,214 ms, *SD* = 384 ms) was significantly faster than from the top views ϕ_0=45° (*M* = 2,311 ms, *SD* = 411 ms), *t*(45) = −3.606, *p* = .002, and from the bottom views ϕ_2=135° (*M* = 2,361 ms, *SD* = 414 ms), *t*(45) = −5.383, *p* < .001. Thus, both accuracy and RT indicated that the middle views were easier than the top and bottom views. See the plots A and B in the Figure 4.4 for the action recognition accuracies and RTs in all three view angles ϕ, respectively.

There was a marginal interaction effect between angle θ and angle ϕ on accuracy, *F*(8, 360) = 1.953, *p* = 0.051, and a significant interaction effect on RT, *F*(6.306, 283.762) = 3.150, *p* = 0.005. This indicates that angle θ had different effects on accuracy and RT depending on angle ϕ and vice versa. Given the large number of potential comparisons, I did not follow up on this interaction with statistics, but the heatmaps in Figure 4.2(A) and Figure 4.2(B) illustrate the action classification accuracies and RTs of the participants in all fifteen viewpoints, in turn. At middle-height viewpoints (ϕ_1=90°), the angle θ made relatively little difference to accuracy and RT. Within the top and bottom viewpoints, angle θ made more of a difference and it was here that the exact front view was noticeably less efficient.

Additionally, I ran a correlation between the 15 accuracies of the 15 views in Figure 4.2(A) and the 15 RTs of the same 15 views in Figure 4.2(B). The view accuracy and the view RT were strongly negatively correlated, *r*(13) = −.88, *p* < .001. This means action recognition was predictably more accurate and faster (shorter) from some views and it was less accurate and slower (longer) from other views. Therefore, the efficient and inefficient views were congruent for both view accuracies and view RTs.

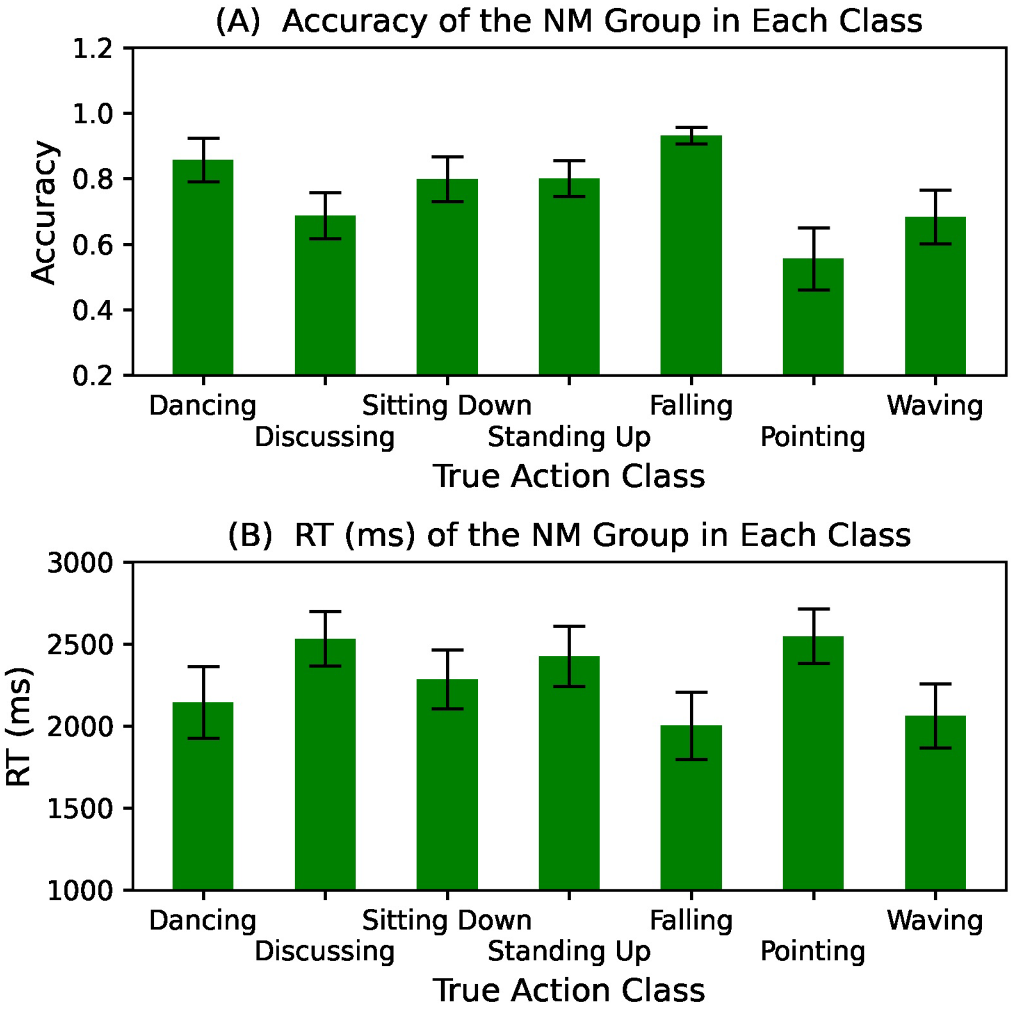


Figure 4.5. The accuracies (A) and the RTs (B) of the human observers in recognizing each class of actions.

4.3.3 Efficient and Inefficient Action Classes

A major aim of this study was to identify the efficient and inefficient views for action recognition. However, I also calculated some descriptive statistics to highlight the efficient and inefficient action classes (or categories). Efficient action categories are recognised by the observers with higher accuracy and shorter RT. However, inefficient action categories are recognised with lower accuracy and longer RT. I did not compute any inferential statistics to compare the recognition means of the classes. The reason was that this was not the real aim of this study and each participant did not do the task with all 7 classes, but

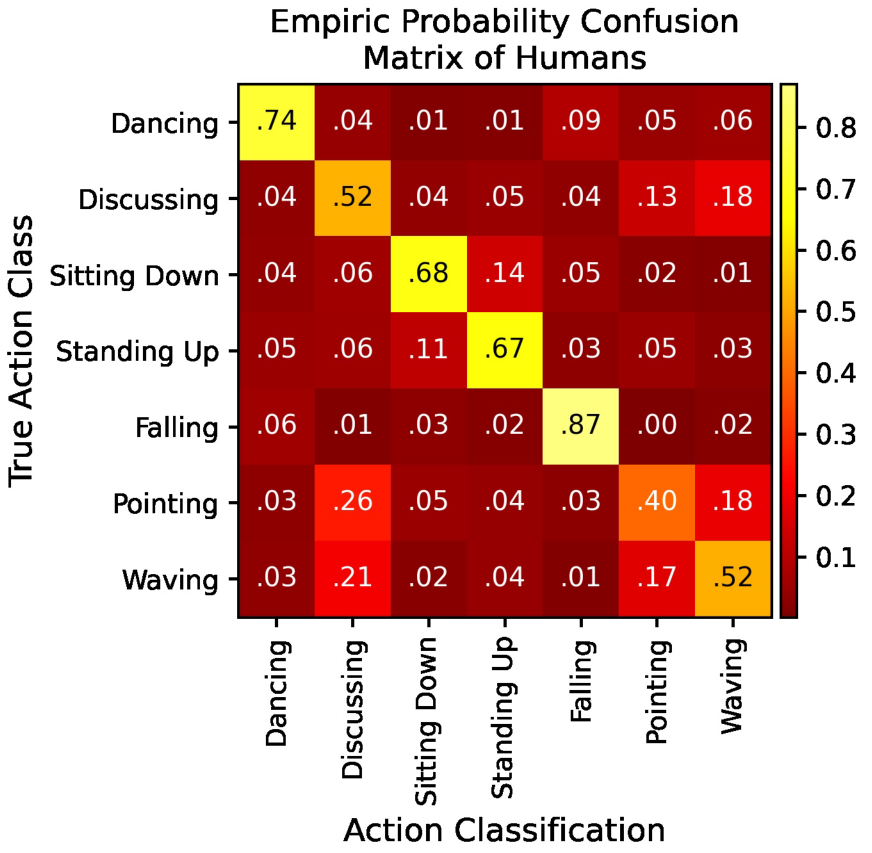


Figure 4.6. The confusion matrix from the action classifications of the human participants.

they saw and classified only 4 random classes out of 7.

The bar chart in the Figure 4.5(A) shows that action recognition was more accurate for dancing (*M* = .858, *SD* = .157), sitting down (*M* = .798, *SD* = .178), standing up (*M* = .801, *SD* = .145) and falling down (*M* = .932, *SD* = .064) than for discussing (*M* = .688, *SD* = .171), pointing (*M* = .556, *SD* = .246) and waving (*M* = .683, *SD* = .185). However, the chart in the Figure 4.5(B) displays that dancing (*M* = 2,146 ms, *SD* = 517 ms), falling down (*M* = 2,003 ms, *SD* = 518 ms) and waving (*M* = 2,062 ms, *SD* = 442 ms) were more quickly recognised than discussing (*M* = 2,534 ms, *SD* = 402 ms), sitting down (*M* = 2,287 ms, *SD* = 466 ms), standing up (*M* = 2,426 ms, *SD* = 490 ms) and pointing (*M* = 2,550 ms, *SD* = 425 ms). There is not a lot of coherence because some action may have happened earlier in some videos, and it may have happened later in some other videos. For this reason, I suggest that the class accuracy is more genuine than the RT as a measure of the class performance. Therefore, the efficient action classes for action recognition were dancing, sitting down, standing up and falling up, while the inefficient classes were discussing, pointing and waving. This may have happened because participants tented to slightly confuse discussing, pointing and waving as shown in the confusion matrix in the Figure 4.6.

4.4 Conclusions

The first goal of this study was to examine whether the recognition of the action classes in the MVVHAs is viewpoint-dependent. The results showed that both accuracy and RT was significantly affected by the viewpoints. Furthermore, the follow-up results also showed that both changes in the azimuth angle θ and the elevation angle ϕ of the viewpoints caused alone significant effects on both action recognition accuracy and RT. Therefore, the recognition of MVVHAs is viewpoint-dependent. This may have been because the high level of similarities between the action classes in MVVHAs which was highlighted by the confusion matrix. This is in line with the studies showing the relation between object similarities and the viewpoint dependency of object recognition (Hayward & Williams, 2000; Leek & Johnston, 2006; Tarr & Hayward, 2017; Tarr & Pinker, 1990; Tjan, 2001).

The second aim of this chapter was to discover the efficient and inefficient viewpoints for humans in action classification. With respect to the angle θ, the side-front (right-front) viewpoints were the most efficient because the action classification of the human participants was the most accurate and the fastest from them. Interestingly, the exact front views were not the most efficient. The back views were most inefficient given that the humans performed the least accuracies and the longest RTs when they classified the actions from these back views. Concerning the angle ϕ, the middle views were the most efficient because human participants scored the highest accuracies and the shortest RTs from these middle view positions. On the other hand, the top and the bottom views were inefficient as from them the action recognition accuracy and the RT of the humans were the lowest and the longest from these views, respectively.

These results will be vital in the chapters 5 and 6. In chapter 5, I will look at whether the efficient and inefficient views for human and robotic observers in action classification are similar. In chapter 6, I will benchmark the active action classification of humans based on how often they select the efficient and inefficient views.

5 Efficient and Inefficient Views for the Action Recognition of Robotic Observers

5.1 Introduction

There are different ways to show whether some active vision models select views efficiently for action classification. I discuss here the most two obvious ones. One way is testing whether the action classification of the active models is more accurate than the baseline models which are non-moving, randomly moving and constantly moving in the same direction. Another way is highlighting whether the active models select more often the efficient views rather than the inefficient ones. To verify the second evidence, we need to know first which views are efficient and which ones are inefficient. Therefore, in this chapter, I mainly investigate the efficient and inefficient views for passive computer vision models.

I defined efficient and inefficient views for humans in chapter 4. The efficient views for humans were the views from where their accuracies were higher than their averaged accuracy from all views and from where their RTs were shorter (faster) than their averaged RT from all views. On the other hand, the inefficient view for humans were the views from where their accuracies were lower and their RTs were longer (slower) than their averaged accuracy and their averaged RT from all views, respectively. However, in this chapter, I define the efficient and inefficient views for computer vision models. The efficient views for robotic observers are the views from where the accuracies of the robotic classifications are higher and their classification losses are lower than their own averaged accuracy and loss from all views, in turn. Instead, I define the inefficient views for the robots as the views from where their accuracies are lower and their losses are higher than their corresponding averages performed from all viewpoints.

To discover the efficient and inefficient views for robotic observers, I developed some basic computer vision models and trained them to classify human actions in videos with different view positions. For this computer vision study, I used the videos of a larger version (r.3.5) of my MVVHA dataset, which has 40 viewpoints (8 angles θ x 5 angles φ). Then, I computed the accuracies and the losses of the models performed from each view and compared these performances of different views.

CNNs have been the most efficient models for computer vision tasks. Since the 2d CNN named AlexNet (Krizhevsky et al., 2017) extensively outperformed the traditional methods to classify single objects in single images, many other researchers designed numerous other 2d CNNs that were very accurate for object classification (He, Zhang, Ren, & Sun, 2016; Simonyan & Zisserman, 2014; Szegedy et al., 2015) and even for real-time object detection (Redmon et al., 2016) on single images. According to Yamins and DiCarlo (2016), 2d CNNs are very accurate in analysing single images because, like the brain's visual system, they can identify both the local features in the shallower and global features in the deeper layers. Local features like edges, angles, surfaces and colours are detected by the shallower convolutional layers that process the patches of an image individually without the surrounding noise. However, the global features such as object classes and action classes are uncovered by the deeper fully-connected layers that integrate all local features coming from all patches of the whole image into fewer global features. In fact, the visual receptive field of each neuron in CNNs increases from shallower to deeper layers as well as the neuron receptive field is wider and wider for deeper and deeper visual areas of the brain's visual system. Similarly, the retinotopy of each layer declines in deeper and deeper layers in CNNs so does the retinotopy for deeper and deeper visual areas of the brain's visual system.

However, 2d CNNs were designed to only extract the spatial features of individual images and cannot extract the spatiotemporal features of videos which are the changes of spatial features over time. Computer vision models can significantly benefit from spatiotemporal features to analyse videos. One reason is that their prediction would be more robust to some noisy frames. Another reason is that they would be able to recognise patterns of changes such as actions that develop over time. 2d CNNs with temporal pooling layers (Karpathy et al., 2014), 3d CNNs (Ji et al., 2013; Tran et al., 2015) and RCNN (Donahue et al., 2015; Kubilius et al., 2018; Li et al., 2018; Liao & Poggio, 2016; Ng et al., 2015) are some more advanced types of CNNs that can extract and exploit the spatiotemporal features. See chapter 1 for a more detailed description of these different types of CNNs.

However, the robotic observers in the study of this chapter investigating the efficient and inefficient viewpoints for the action recognition of robotic observers were RCNNs and 2d CNNs. I chose RCNNs over the non-recurrent CNNs (2d CNNs, 2d CNNs with temporal pooling layers and 3d CNNs) because RCNNs are more appropriate than the non-recurrent CNNs for active action recognition which was addressed in chapter 7. I also utilised non-recurrent 2d CNNs in this chapter because they are an inner component of RCNNs. The RCNNs are more suitable for active action recognition because, as recurrent models, RCNNs can exploit spatiotemporal features and make a prediction at each time point. Thus, they can predict the best next viewpoint movement at each timepoint and this prediction would be dependent on the observation of the previous timepoints by spatiotemporal features. 2d CNNs can make a prediction at each timepoint, but its predictions are independent of the previous observations with only spatial features. 2d CNNs with temporal pooling layers and 3d CNNs can also exploit spatiotemporal features, but they can only make one prediction per video rather than one prediction per image (timepoint) as they analyse a video as a whole. Additionally, I chose the RCNNs over the non-recurrent CNNs because of many other advantages of RCNNs over the non-recurrent CNNs that are detailed in chapter 1.

Hence, I utilized two types (groups) of models which could either be non-recurrent 2d CNNs or RCNNs. The non-recurrent 2d CNNs were 18-layer ResNets (He, Zhang, Ren, & Sun, 2016) which were trained on ImageNet (Russakovsky et al., 2015) for object classification and subsequently finetuned by me on the video images of my MVVHA for actions classification. These non-recurrent 2d CNNs classified each image individually and independently from the other images of the same video. However, following a popular recent suggestion (Donahue et al., 2015; Ng et al., 2015), I designed a RCNN architecture with two serial components, a 2d CNN feature extractor and a recurrent 1d layer. I chose the finetuned 18-layer ResNet without the last fully-connected layer as the feature extractor and the long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) as the recurrent layer.

I had several predictions for the computer vision study I present in this chapter. First, the overall accuracy of both non-recurrent and recurrent models would be relatively low. I intentionally lowered their overall performance by reducing the amount of training videos. Some previous pilot trials showed that non-recurrent models with NMs could perform accuracies of about 0.99 when they were trained on all videos of 35 actors out of 65. If the action recognition accuracy of NM non-recurrent models were already nearly 1, which is the maximum possible value, there would not have been any extra room for improvement and highlighting significant differences in accuracies between different conditions would have been impossible. In this way, showing significant differences in accuracy between different viewpoints would have been impossible. Likewise, exposing significant gains in accuracy from the passive to the active models would not have been possible. Therefore, I only included the videos of five actors in the training data to reduce the number of training videos and intentionally decline the models' overall accuracy.

Second, the recurrent classifiers would be more accurate than the non-recurrent ones overall because of the reasons I discussed in chapter 1. In this case, it is mainly because the recurrent models integrate time information rather than the recurrent models. The action classification of recurrent models at the timepoint t is a function of the frames of the timepoints 0 to t, whereas the classification of non-recurrent models at the timepoint t is only a function of the frame at time point t. Therefore, the recurrent models know more about the actions because they can collect multiple observations of the actions at different timepoints.

Third, because of the same reason, the recurrent models would be even more accurate than the non-recurrent models in classifying the actions in the video frames with larger timepoint t. In fact, the recurrent models would collect more (t + 1) observations and would have more evidence about the actions for larger t. Meanwhile, since the non-recurrent models process each frame individually and independently, the non-recurrent models can only have one observation and would have less evidence available regardless of the timepoint t.

Fourth, the recurrent models would be more accurate than the non-recurrent ones in classifying the time-inversed actions. Actions such as sitting down (standing-to-sitting) and standing up (sitting-to-standing) are inverted in time to each other. The recurrent models should have a better representation of the direction of the movements because they observe the development of the action in time. Instead, the direction of the movements in time-inversed actions should be less clear to non-recurrent models given that they classify each image individually and independently from the previous frames of the same videos.

Fifth, I expected the pattern of efficient and inefficient views for the computer models would be similar as for the humans which I spotted in the previous chapter 4. If this is true, the view efficiency of humans and the view efficiency of computer models are expected to be linearly correlated. Specifically, based on how I defined the view efficiency for humans and robots in terms of view accuracy, view RT and loss, the view accuracies of humans should be positively correlated to the view accuracies of the robotic observers and negatively correlated to the view losses of the same models. On the other hand, the human RTs should be negatively correlated to the view accuracies of the classifiers and positively correlated to the view losses of the models. Because the view accuracy and the view losses of the same group of models were highly correlated, $r \geq -.97$, I will only address whether the human view accuracies and the human view RTs are correlated to the robotic view accuracies. Any correlation between the view efficiency for humans and the view efficiency for the classifying models would be evidence of two different conclusions which are not mutually exclusive. One, the computer vision models can approximate the complicated function of the biological visual system recognising actions. Two, the efficient views objectively provide more visual information about the action than the inefficient ones, independently of the type of the observers.

5.2 Methods

5.2.1 Dataset MVVHA

I fed the robotic vision models with the images of the version r.3.5 of my own dataset MVVHA. This dataset version has 24,570 MVVs that were generated by combining 7 classes, 27 VR animation per class, 65 actors and 2 mirrors. Each MVV has 40 SVVs with 40 different viewpoints. Therefore, there are 982,800 SVVs (24,570 MVVs x 40 viewpoints). The 40 views were defined by combining 8 angles θ and 5 angles ϕ. The 8 azimuth angles θ were: −180° (back), −135° (right-back), −90° (right), −45° (right-front), 0° (front), +45° (left-front), +90° (left), +135° (left-back). The 5 different elevation angles ϕ were: 0° (top), +45° (middle-top), +90° (middle), +135° (middle-bottom), +180° (bottom). Every SVV has 60 frames for 2 seconds, making a refresh rate of 30 Hz and each frame was 224-pixels shared encoded in RGB colour channels.

5.2.2 Computer vision models

In this study, I trained two types of computer vision models to classify the actions in the video images in MVVHA. The models could either be recurrent or non-recurrent classifiers. Instead of training, validating and testing a single model per model type, I trained, validate and tested ten models per models type to increase the statistical power of the results. Thus, I trained ten non-recurrent classifiers and ten recurrent classifiers. I made ten different splits of the dataset into training, validation and test data for each group of models, each split per model. Thus, each model within the same model type was trained, validate and tested with a different split of training, validation and test data. However, all groups of models were trained, validated and tested with the same ten splits.

The non-recurrent classifiers classify each image of a video individually and independently from the other images of the sequence. The non-recurrent classifiers were 18-layer ResNets (He, Zhang, Ren, & Sun, 2016), deep neural networks with convolutional and fully-connected layers. Firstly, I took the standard 18-layer ResNet which is pretrained to classify the 1000 classes of objects on the ImageNet dataset (Russakovsky et al., 2015). Next, I replaced the last fully-connected layer with size 1,000 with another one with size 7 to classify the seven action classes in my dataset. Lastly, I finetuned (re-trained) ten instances of the network with different training data for more statistical power of the findings.

The recurrent classifiers are deep neural networks designed to classify a video by integrating the time information of its image sequence. Each of their classification of an image is dependent on (or is a function of) the images of the previous time points and the current time point. The architecture of the recurrent classifiers can be described with two serial components:

1. A non-recurrent feature extractor which was a non-recurrent classifier (finetuned 18-layer ResNet) without the last layer. This extracts the features of the video frames, individually and independently. Therefore, this takes a batch of videos as input with data size T x B x 3 x 224 x 224 where T is the number of frames per video, B is the number of videos per batch and 3 x 224 x 224 is the shape of each image. This component outputs features with size T x B x 512;
2. A recurrent feature classifier which further includes 2 layers:
    - 2.1. A long short-term memory (LSTM) layer that integrates the image features across time (Hochreiter & Schmidhuber, 1997; Sak et al., 2014). The size of its hidden state was set to 500 such that it takes a batch of video features with shape T x B x 512 as input and outputs a batch of hidden states with shape T x B x 500;
    - 2.2. A feature classifier which is a fully-connected layer that individually classifies the hidden states of the of all time points with size T x B x 500. This layer has size 7 as much as the number of action classes in my dataset. In this way, it returns the class values with shape T x B x 7.

5.2.3 Training, Validation and Test

I trained all models to classify the actions in the videos. I used all SVVs and all views of the computer vision dataset r.3.5 to train, validate and test all models. However, I only took 6 frames out of 60 per video and given that each video was 2-second long, all models saw the videos with a refresh rate of 3 Hz. The 6 frames per video were [0, 10, 20, 30, 40, 50] plus a random integer t in the range $0 \le t < 10$. For example, if the random integer for a video was 3, then the six frames were [3, 13, 23, 33, 43, 53]. The loss function was the cross entropy loss (Mao et al., 2023; Zhang & Sabuncu, 2018) for both non-recurrent classifier and recurrent classifier. I trained, validated and tested all computer vision models with python code with some python libraries. The most relevant libraries were PyTorch (Paszke et al., 2019), NumPy (Harris et al., 2020) and CalaPy (pypi.org/project/calapy). I personally coded CalaPy as a whole.

I trained the recurrent classifiers in three steps: non-recurrent feature extractor finetuning, feature extraction, and recurrent feature classifier training. The non-recurrent feature extractor finetuning was the finetuning of the ten non-recurrent classifiers of this study. In feature extraction, I extracted the image features of whole dataset ten times, each time with one of my ten non-recurrent feature extractors which were the trained non-recurrent classifiers (finetuned ResNets) without the last layer. This produced ten datasets of features. In recurrent feature classifier training, I trained each of the ten recurrent feature classifiers with the training data of a different feature dataset.

I ran all experiment steps on the 24 high-performance NVidia GPU cards of the Ceres cluster. The GPU models were GTX1080Ti and RTX2080. The Ceres cluster is a computational cluster built using the Rocks Clustering Solution with CentOS Linux. For computational purposes, the cluster has 1096 processing cores (2192 with hyperthreading) provided by servers with a mix of Intel E5-2698, Intel Gold 5115, 6152 & 6238L processors, and between 500Gb & 6Tb RAM each. Storage is provided by a set of storage nodes providing 660Tb of storage. Inter-node connectivity is via 10GbE switches. There are also 24 NVidia GTX & RTX Series GPU cards (16 x GTX1080Ti & 8 x RTX2080) attached via dedicated GPU servers for research purposes.

5.2.4 Multiple dataset splits

The difference between the ten non-recurrent classifiers and between the ten recurrent classifiers was only their finetuning/training experience. They were trained with different training data to increase the statistical power of the findings. I made ten different splits of the dataset version r.3.5 into 3 groups of data: training data, validation data and test data. Every split was done by actors. In each split, I randomly split all 65 actors into 5 training actors, 10 validation actors and 50 test actors, and I flagged all images of the training actors as the training data, all images of the validation actors as the validation data and all images of the test actors as the test data. Any actor could not fall in the group of training actors in more than one split, such that no single image was fed to more than one non-recurrent classifier and one recurrent classifier during their training. Furthermore, no single actor was seen by more than one non-recurrent classifier and one recurrent classifier for their training.

I used the same identical ten splits of actors for the non-recurrent classifiers and the recurrent classifiers. Thus, any unknown variables are controlled because they are trained, validated, and tested with the same identical data. This makes the results of the two types of models more comparable.

5.2.5 Design

The design was 8 x 5 within-subject (within-models) for the computer vision experiment. The angle θ and the angle ϕ were the two independent variables. There were 8 angles θ: −180° (back), −135° (right-back), −90° (right), −45° (right-front), 0° (front), +45° (left-front), +90° (left), +135° (left-back). I manipulated the angle ϕ with 5 conditions: 0° (top), +45° (middle-top), +90° (Middle), +135° (middle-bottom), +180° (bottom). All conditions of the two independent variables (thetas and phis) of the human study of chapter 4 are included in the computer vision study of the chapter. These matches of views of the human and robot studies made their results about efficient and inefficient views comparable between the humans and robots of the two studies.

I analysed two dependent variables: action classification accuracy and action classification loss.

5.3 Results

5.3.1 Non-Recurrent vs Recurrent

Generally speaking, the recurrent models were more accurate (M = 0.69, SD = 0.04) in action classification than the non-

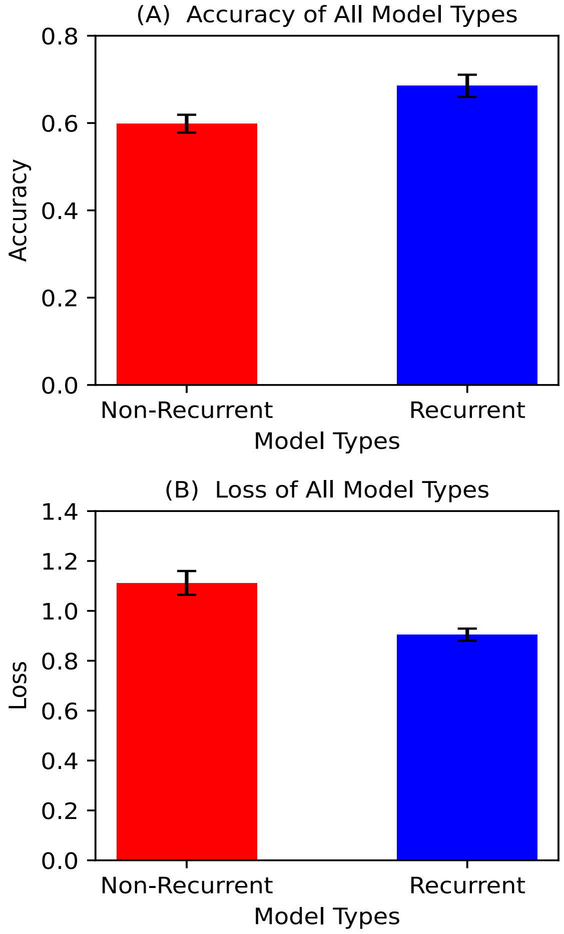
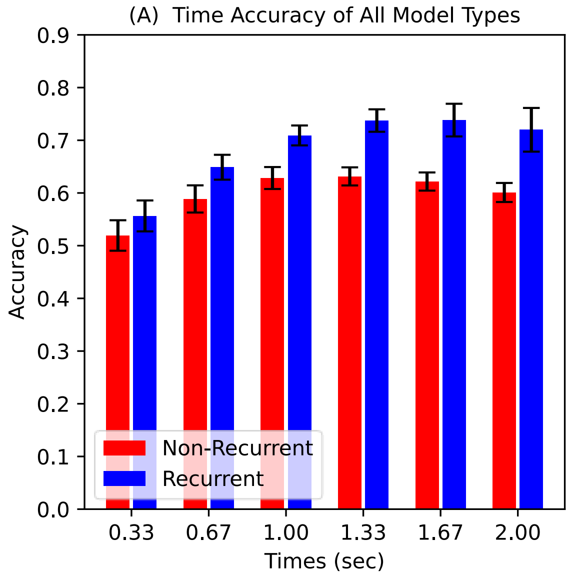


Figure 5.1. Classification accuracy (A) and classification loss (B) of the Non-Recurrent and Recurrent models.

recurrent models (*M* = 0.60, *SD* = 0.03). Figure 5.1 shows the overall accuracy and the loss of both groups of models. This result suggests that the time integration of the recurrent models improves the video analysis for action classification.



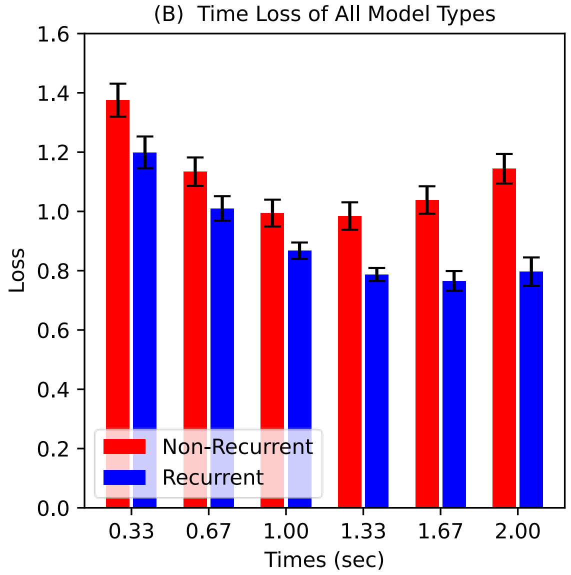


Figure 5.2. The classification accuracy (A) and the classification loss (B) of both non-recurrent and recurrent classifiers in all 6 used time points. They were averaged across all SVVs.

The last statement is even more obvious in Figure 5.2(A). The recurrent classifiers outperformed the non-recurrent ones in all timepoints of the SVVs. Importantly, the difference in accuracy of the types of models gets larger with the number of observations (time points). That is because the recurrent models can integrate the time information and accumulate the frames (or observations) over time whereas the non-recurrent models only process each frame individually and independently from the previous observations. The non-recurrent classifiers scored the lowest accuracy at the first time point (*M* = 0.52, *SD* = 0.04). Next, their accuracy constantly raised and stabilised from the third time point to the last one in a range between 0.60 and 0.63. It peaked at the fourth time point (*M* = 0.63, *SD* = 0.02) and, following that, it slightly decreased reaching a low at the sixth and last time point (*M* = 0.60, *SD* = 0.03). Likewise, the classification accuracy was the lowest at the first observation (*M* = 0.56, *SD* = 0.04). After, it steadily increased with the number of observations, and it also stabilized from the third time point to end in a range from 0.71 to 0.74. The peak was at fifth time point (*M* = 0.74, *SD* = 0.04) and then, it slightly declined at the last time point (*M* = 0.72, *SD* = 0.06). Both recurrent and non-recurrent models peaked at the fourth and fifth time points of the videos which are the time points 3 and 4 (the 6 time points were numbered from 0 to 5). Therefore, I chose to only show the fourth time point (time point 3) of the next results when their accuracy was approximately the highest.

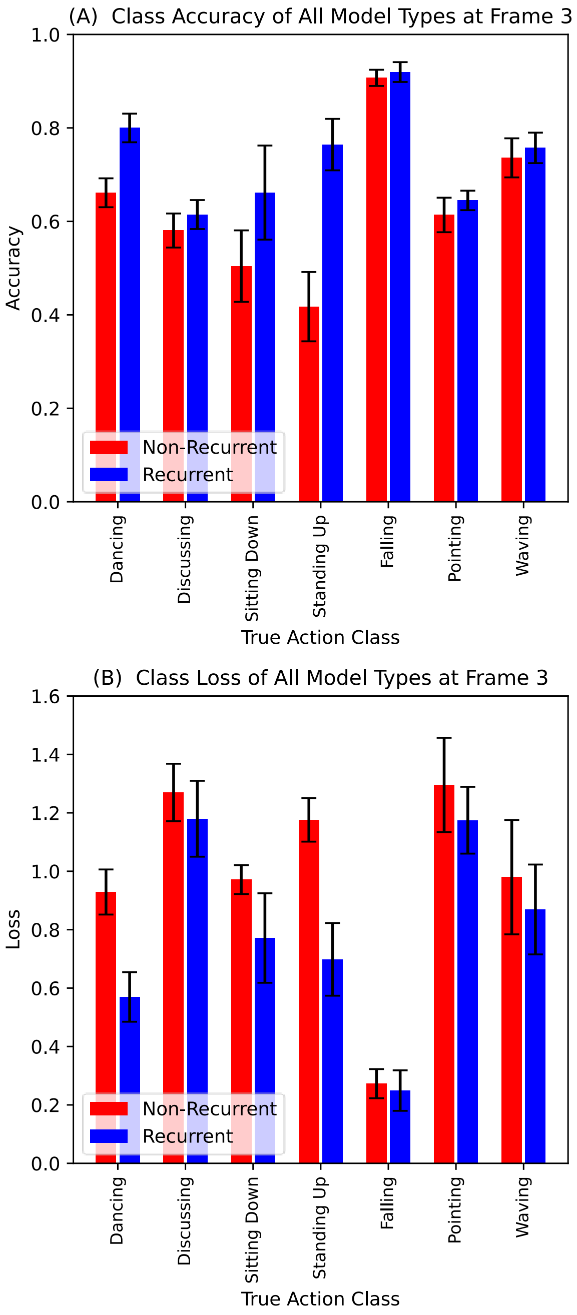
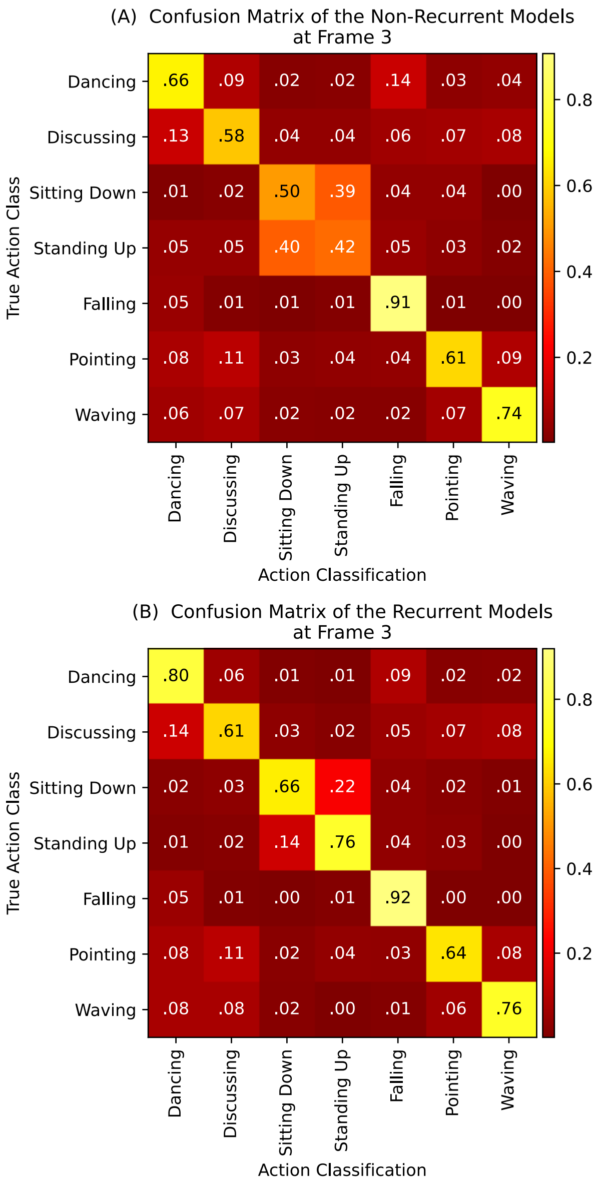


Figure 5.3. The accuracy (A) and loss (B) of the non-recurrent and the recurrent classifiers in classifying each class of actions.

5.3.2 Action Classification of Each Action Class

In Figure 5.3, I individually plotted the accuracy (A) and the losses (B) of the two types of models for each class of actions. Each action class was recognised more accurately by the recurrent models than by the non-recurrent ones. This shows that the addition of recurrent layers which accumulates information through time, improves action classification. The recurrent models outperformed the non-recurrent models with the larger difference for dancing, sitting down and standing up than for the discussing, falling down, pointing and waving. This suggests that discussing, falling, pointing and waving can easily be recognised with a single frame, whereas dancing, sitting down and standing up require more time integration to be correctly categorised. This is particular evident in case of sitting down and standing up which cannot be distinguished by a single frame because they are the inverse of the other with respect to time. In this case of inverse actions, time integration is a requirement. Moreover, falling was the most recognised action class by both non-recurrent (*M* = 0.91, *SD* = 0.02) and recurrent (*M* = 0.92, *SD* = 0.03) classifiers.



(A) Confusion Matrix of the Non-Recurrent Models at Frame 3

	Dancing	Discussing	Sitting Down	Standing Up	Falling	Pointing	Waving
Dancing	.66	.09	.02	.02	.14	.03	.04
Discussing	.13	.58	.04	.04	.06	.07	.08
Sitting Down	.01	.02	.50	.39	.04	.04	.00
Standing Up	.05	.05	.40	.42	.05	.03	.02
Falling	.05	.01	.01	.01	.91	.01	.00
Pointing	.08	.11	.03	.04	.04	.61	.09
Waving	.06	.07	.02	.02	.02	.07	.74

(B) Confusion Matrix of the Recurrent Models at Frame 3

	Dancing	Discussing	Sitting Down	Standing Up	Falling	Pointing	Waving
Dancing	.80	.06	.01	.01	.09	.02	.02
Discussing	.14	.61	.03	.02	.05	.07	.08
Sitting Down	.02	.03	.66	.22	.04	.02	.01
Standing Up	.01	.02	.14	.76	.04	.03	.00
Falling	.05	.01	.00	.01	.92	.00	.00
Pointing	.08	.11	.02	.04	.03	.64	.08
Waving	.08	.08	.02	.00	.01	.06	.76

Figure 5.4. The confusion matrices of both non-recurrent (A) and recurrent (B) models.

Figure 5.4 displays the empiric probability confusion matrices of the non-recurrent (A) and the recurrent (B) models. Overall, the recurrent models returned more often the correct class and less often the wrong classes than the non-recurrent classifiers. The non-recurrent classifiers tend to largely confuse sitting down and standing up. That is mainly because these actions are time-inversed and the non-recurrent models which does not integrate information across time are not able to distinguish them. However, these confusions are overtaken with the time integration of the recurrent classifiers. This pattern shows again that classification of dancing, sitting down and standing up benefit more from recurrent models. Instead, discussing, falling down, pointing and waving can easily be classified with a single frame as the simple non-recurrent classifiers tend to not confuse them.

5.3.3 Efficient and Inefficient Views for Action Classification

The Figure 5.5(A) shows how the action classification accuracy of both model types in the time point 3 changed as function of the viewpoint angle θ. Overall, the recurrent classifiers were more accurate than the non-recurrent classifiers in all θ conditions. However, for both models the left and the right views are approximately symmetric. The main explanation is that the MVVs of the non-mirrored 3D animations and the MVVs of the mirrored 3D animations and therefore, the information showed in the left views and right views were balanced across all MVVs. The accuracy of both type of models had two major lows at the back and front views with θ=−180° and θ=0°, respectively. The accuracy was higher in both model types in the side views which are the right-back, right, right-front, left-front, left, left-back views. In turn, these were the views with θ=−135°, θ=−90°, θ=−45°, θ=+45°, θ=+90° and θ=+135°. Accuracy from the back views with θ=−180 were the lowest for both the non-recurrent (*M* = 0.59, *SD* = 0.03) and recurrent (*M* = 0.68, *SD* = 0.03) classifiers. Then, it raised reaching a high at right views with θ=−90 for all non-recurrent (*M* = 0.65,

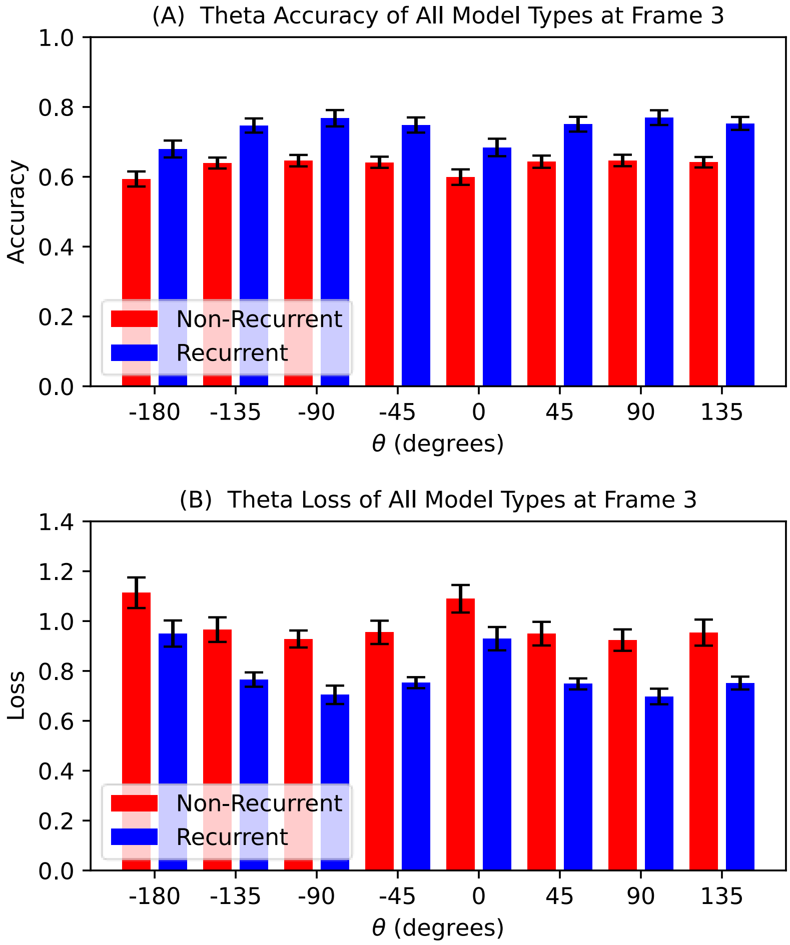


Figure 5.5. Action classification accuracy (A) and loss (B) of both model types in each viewpoint angle θ in the time point (video frame) 3.

SD = 0.02) and recurrent (*M* = 0.77, *SD* = 0.03) models. After that, it diminished making a low in the front views with θ=0° for both non-recurrent (*M* = 0.60, *SD* = 0.03) and recurrent (*M* = 0.68, *SD* = 0.04) classifiers. Next, it started increasing again printing a new high at the left views with θ=+90° for the non-recurrent (*M* = 0.65, *SD* = 0.02) and recurrent (*M* = 0.77, *SD* = 0.03) classifiers. Finally, it dropped once again as we reapproach the back views with θ=±180°. This pattern shows that the efficient viewpoints for the computer vision models were the sided one while their inefficient views were the back and font ones. At the same time, the front views are slightly more efficient than the back views.

Let us now talk about how the models performed in the 5 conditions of the viewpoint angle

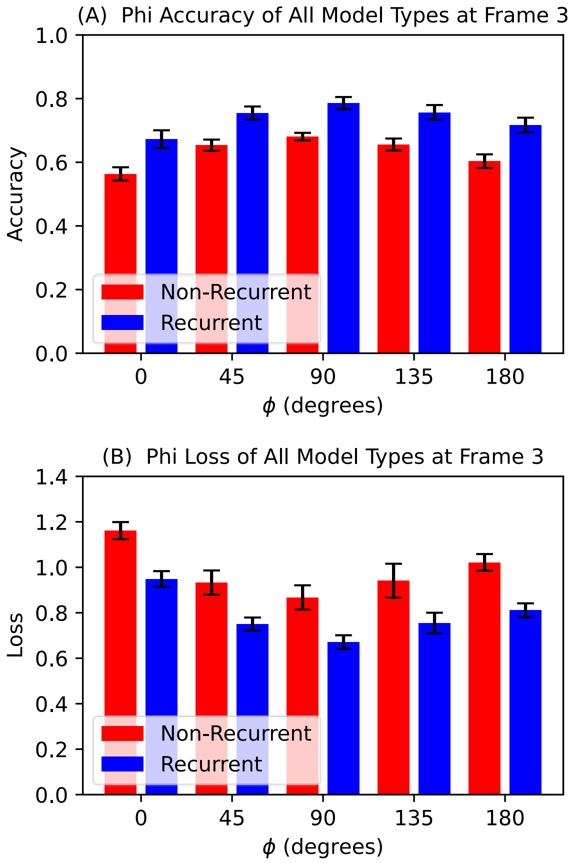


Figure 5.6. Action classification accuracy (A) and loss (B) of both non-recurrent and recurrent classifiers in each viewpoint angle ϕ in the time point 3.

ϕ. The results are visualized in Figure 5.6. Overall, all the recurrent classifiers outperformed the non-recurrent models in each angle ϕ. However, the action classification accuracy in the middle views with ϕ=90° was the highest for both non-recurrent (*M* = 0.68, *SD* = 0.02) and recurrent (*M* = 0.79, *SD* = 0.03) models. Next, their accuracy gets lower as the views get closer to either the very top with ϕ=0° or to the very bottom with ϕ=180°. Thus, the very top and the very bottom views were the lows of their accuracy, meanwhile the lower viewpoints were slightly better than the top ones. The lowest accuracy was in the very top views for both non-recurrent (*M* = 0.56, *SD* = 0.03) and recurrent (*M* = 0.67, *SD* = 0.04) models. Taking together these results about the ϕ accuracy of both model type, the most efficient views are the middle ones which are followed by the middle-bottom and the middle-top, in turns. On the other hand, the inefficient views were the very top and the very bottom views.

Figure 5.7 highlights the view accuracy of the non-recurrent and of the recurrent classifier. it shows four major results. First, the recurrent models beat the non-recurrent one in each of all 40 views. Second, their performances were symmetric for the left and right views with θ=±45°, θ=±90° and θ=±135°. Third, both types of models scored poorly in the top views

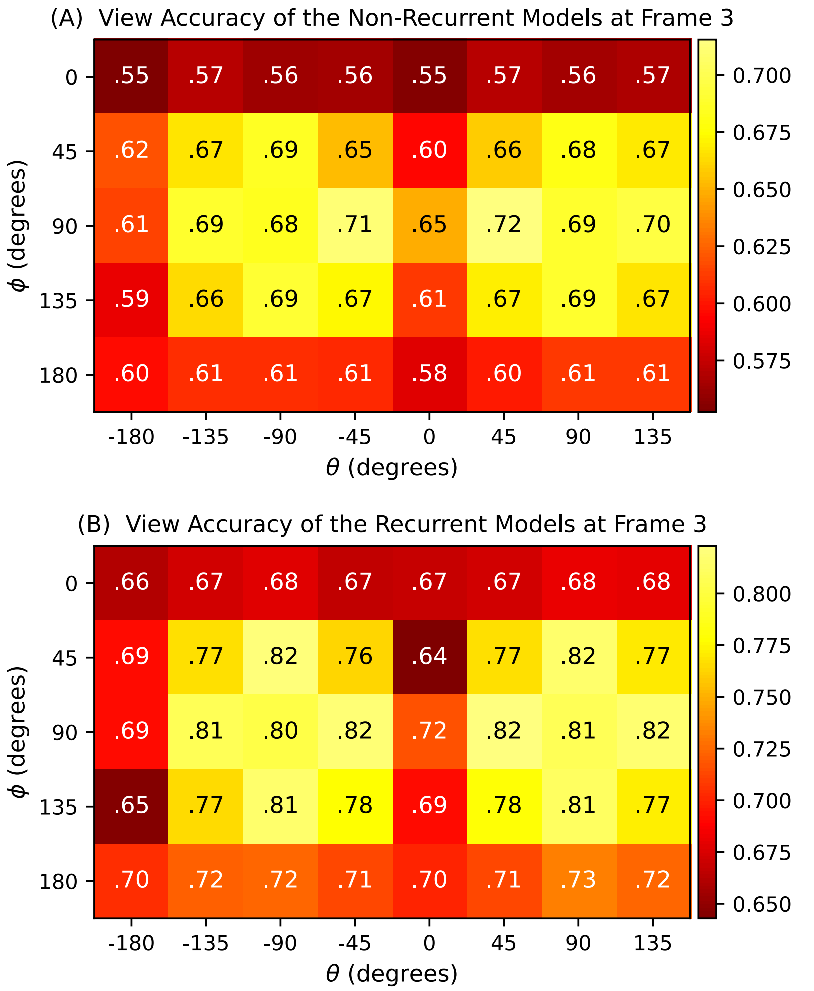


Figure 5.7. The accuracy of both (A) non-recurrent and (B) recurrent models in all 40 views.

with ϕ=0°, and in the bottom views with ϕ=180°. Fourth, their accuracy tended to be lower in the back views with θ=−180°, and in the front views with θ=0°. Therefore, the most efficient views for both models were the side and mid-height views with θ=±135°, θ=±90°, θ=±45° and with ϕ=+45°, ϕ=+90° ϕ=+135°. The inefficient views were the top, bottom, front and back views with θ=±180°, θ=0° and with ϕ=0° ϕ=+180°.

The efficient and inefficient views for the non-recurrent computer models were mostly the same as for the recurrent models. If this statement is true, we would expect the non-recurrent view accuracies in the heatmap (A) of Figure 5.7 are positively correlated to the recurrent view accuracies in the heatmap (B) of the same figure. In fact, the accuracies of the non-recurrent models should be lower in same views as of the recurrent models. On the other hand, the accuracies of the non-recurrent and recurrent models should higher in same other views. Therefore, I computed a Pearson correlation coefficient to assess the linear relationship between the non-recurrent view accuracies and the recurrent view accuracies. Each datapoint of this correlation corresponds to the non-recurrent accuracy and recurrent accuracy of one viewpoint. Given that there were 40 views in my experiment, the datapoints of this correlation were also 40. The correlation matrix between the view performances of my observers in Figure 5.8 includes the results of the correlation between the view performances of both model types. There was a strong positive correlation between non-recurrent and recurrent view accuracies, *r*(38) = .95, *p* < .001. This is very strong evidence that the efficient and inefficient views were the same for both types of models. This may be because of two main reasons. One, the two types of models are equivalent as the recurrent models include the non-recurrent models without their own last classifying layer as feature extractors. Two, the efficient and inefficient views for both types of models are objectively efficient and inefficient views: there is objectively more visual information about the actions from the efficient views than from the inefficient ones. Therefore, even if I change the model types, I still get the same efficient and inefficient views.

Interestingly, the efficiency of the views for both groups of models were also correlated with the efficiency of the views for the human participants on the experiment with no view motion in chapter 4. A view for the models is efficient if their accuracy is higher and their loss is lower from that view than average of all views. Instead, it is inefficient for the robotic observers if their accuracy is lower and their loss is higher in that view than the average of all viewpoints. However, a view for humans is efficient if their classification accuracy is

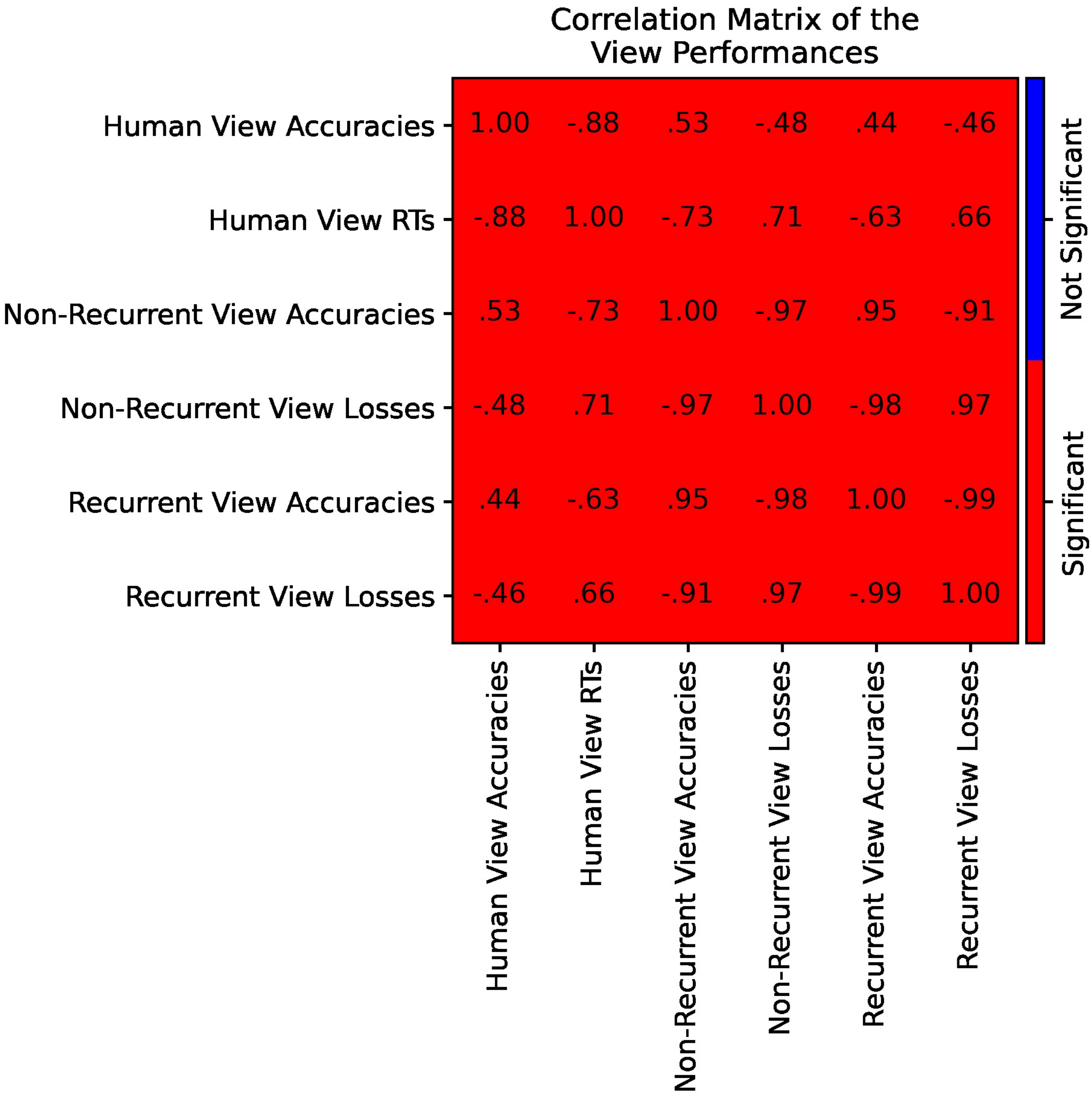


Figure 5.8. Correlation matrix between the different view performances of the distinct types of observers.

higher and their RT is shorter (faster) in that view than average of all views. Otherwise, if their accuracy is lower and the RT is longer (slower) in one view than the average of all views, that view is inefficient for human observers. Therefore, the view accuracies of both non-recurrent and recurrent models in heatmaps (A) and (B) of Figure 5.7 should positively correlated to the view accuracies and negatively correlated to the view RTs of the human participants in the study of chapter 4 (see Figure 4.2). I calculated 4 Pearson correlation coefficients to instigate the linear relationship between the non-recurrent view classification accuracies and the human view classification accuracies, between the recurrent view accuracies and the human view accuracies, between the non-recurrent view accuracies and the human view RTs, and the recurrent view accuracy and the human view RTs. There was an issue in calculating these 4 correlations because each data point corresponds to one single view and there were 40 views in the computer vision experiment while there were only 15 views in the psychology experiment of chapter 4. However, all views of the human experiment were included in the computer vision experiment even if some views of the computer vision design were not in the human design. Specifically, the human experiment did not have the very top and the very bottom views with ϕ=0° and ϕ=180°. Additionally, I did not include all left views with θ=+45°, θ=+90° and θ=135°. Since I assumed that the left and the right views would be symmetric, filled out the missing left view accuracies and RTs of the humans with their corresponding right view accuracies and RTs of the same human participants, excluding the very top and very bottom views. For instance, accuracy of humans in the left-front and middle-top view with θ=+45° and ϕ=+45° was set equal to the human accuracy in the right-front and middle-top view with θ=−45° and ϕ=+45°, while the human RT in the left-back and middle-bottom view with θ=+135° and ϕ=+135° was equal to the human RT in the right-back and middle-bottom view with θ=+45° and ϕ=+135° and so on. At the same time, I excluded the very top and the very bottom views of the computer models for these 4 correlations. Thus, I had total of 24 datapoints (24 views = 8 angles θ x 3 angles ϕ) in the 4 correlations between view performance of computer models and the view performance of humans. The correlation matrix of the view performances of my human and robotic observers in Figure 5.8 illustrates the results of these 4 correlations. The human view accuracies were positively correlated to both the non-recurrent view accuracies, *r*(22) = .53, *p* = .008, and the recurrent view accuracies, *r*(22) = .44, *p* = .029. On the other hand, the human view RTs were negatively correlated to both the non-recurrent view accuracies, $r(22) = -.73$, $p < .001$, and the recurrent view accuracies, $r(22) = -.63$, $p = .001$. Consequently, the efficient and inefficient views tended to be the same for human and the robotic observers.

5.4 Conclusions

Recent computer visions models have achieved such high performance that can even compete with human vision. This has been a good step forward for social robots given that physical robots need to understand what people are doing in order to interact with them in a physical environment. However, actions are not clearly visible from all possible viewpoints by the robotic observers. Therefore, robots can recognise more accurately if they see the actions from the efficient views rather than the inefficient ones. For this goal, social robots must learn how to estimate the efficiency of all possible views such that they can move to, and see the actions from, the efficient views rather than the inefficient ones. In chapter 7, I will look at whether my active models efficiently select the views. Specifically, I will show whether our active models score higher accuracy than the baseline models and whether the same active models tend to select more often the efficient views than the inefficient one. To verify the latter statement, I need to firstly revel which views are efficient and which one are inefficient for the robotics observers.

The study in this chapter aimed to highlight the efficient and inefficient views for robotic vision. Thus, I trained and tested non-recurrent and recurrent models to classify the same actions from multiple viewpoints which were defined by the azimuth angle θ and the elevation angle ϕ. As expected, the overall accuracy of both modes was low. It was 0.60 and 0.69 for the non-recurrent and recurrent models, in turn. I intentionally made the models perform poorly by training them with only the videos of five actors. Such low accuracy allowed me to show differences between different conditions.

In line with my prediction and many other studies, the recurrent models were more accurate than the non-recurrent models for action recognition. That is because the recurrent models can detect both the body poses and their changes over time, while the non-recurrent models only spot the poses. Additionally, this difference was larger and larger for larger and larger timepoints within the videos. This is because the recurrent models accumulate more and more information about the actions with more and more observations, while the non-recurrent models do not carry any memory of the previous observations.

The recurrent models significantly outperformed the non-recurrent models to recognise sitting down and standing up. The videos containing these actions have the same spatial features as they include the same body poses. However, they have different spatiotemporal features because the body poses in sitting down and standing up change in the opposite direction over time. Therefore, the recurrent models that can catch the spatiotemporal features could correctly distinguish them, while the non-recurrent models that can only extract spatial features tended to confuse them.

With respect to the angle θ, the efficient views were the sided views, the views on the sides of the actors: right-back, right, right-front, left-front, left, left-back. The back and front views were inefficient. With respect to the angle ϕ, the middle, the middle-bottom and the middle-top views are efficient, while the very top and the very bottom views were inefficient. These results will allow me to look in chapter 7 at whether the active models will select the efficient views than the inefficient ones.

The view efficiency for the computer models was highly correlated with view efficiency for the humans. Therefore, the efficient and inefficient views tend to be the same for both robotic vision and human vision in recognising actions. This suggests two further conclusions. One, the computer models approximate the action recognition function of the human brain even if there are some characteristics of their architectures that are not biologically plausible. Two, the efficient views objectively provide more visual information about the actor’s action to any observer regardless of the observer type.

6 Active Action Recognition of Human Observers

6.1 Introduction

The preliminary conclusion of the experiment in chapter 3 was that human observers are efficient active action classifiers because of two main results. One, their action recognition performance in the active SCM condition was slightly better than in the baseline RM condition. Two, they selected the efficient views more often than the inefficient views in the SCM trials. Unfortunately, there were two major mythological issues in the study of chapter 3. The first issue was that the realistic actions that the participants classified were presented with very clear images. This may have been the reason why the action classification accuracy in the SCM trials was not significantly higher than in the RM trials because the human observers may correctly classify the realistic actions from any viewpoint in the case of very clear images. The second issue was that the views of the participants could move within both the RM and SCM trials and the assumption by which the efficient views can be detected by looking at the starting views in which the action recognition was more accurate or faster may have been not valid. The reason for its invalidity was the fact that the viewpoint, the independent variable, which changed within trials was not scrupulously controlled across conditions and the accuracy and the RTs, the dependent variables, were theoretically noisy.

The main aim of these experiments in this chapter 6 was to investigate whether the main results of chapter 3 can be replicated even with different methods in materials, designs, and procedures. Hence, the first objective of the two experiments in this chapter was to tackle the first issue of chapter 3, by showing the actions to the participants with unclear (blurred and transparent) images. The unclarity was at different levels for the two experiments: the images were very unclear in the experiment 1 while they were slightly clear in the experiment 2. In brief, the first objective of the two experiments was to answer the following research question: is the active action recognition of humans in the SCM condition significantly more accurate or faster than in the RM condition in the case of unclear images? The second objective of the experiments in this chapter was to address the second issue of the pilot study in chapter 3, by examining whether human observers in the SCM condition select the efficient views more often than the inefficient views even in the case of more valid efficient and inefficient views of chapter 4. The efficient views were more valid in the study of chapter 4 rather than in the study of chapter 3 because the views of the study in chapters 4 were locked and could not move within trials. In this way, the viewpoint, the independent variable, was better controlled across the different viewpoint conditions and the accuracy and the RTs, the dependent variables, were in principle less noisy in chapter 4 than in chapter 3. Thus, the second objective was to answer this second research question: do human participants select the efficient views rather than the inefficient views, by using the efficient and inefficient views of the chapter 4 which are theoretically more valid?

There were several methodological differences between the two experiments of this chapter and the pilot experiment of chapter 3. The first difference was that the two experiments in this chapter were online studies while the pilot study was a lab study. The participants in the two online experiments were tested online from anywhere they liked on their PCs and MACs. Instead, all participants of the pilot study did their test on the same PC of the same lab.

Another important difference between the studies was that the movement type, the independent variable which could either be RM or SCM, was manipulated between subjects in the online experiments while was manipulated within trials in the lab experiment. This was done to reduce the number of trials per participant. Online studies need to be short for human participants because these participants are far away from the experimenter and they can lose focus on the test very quickly.

Another difference aimed to disturb the action recognition accuracy of the human participants. This was the fact that the images of the videos showing the actions were unclear in these online experiments as I have already mentioned above, whilst they were very clear in the pilot study. The images of the online studies were unclear because they were blurred, transparent and presented at a lower refresh rate. The video refresh rate was 5 Hz in both experiments whilst it was 30 Hz in the pilot study. The images of the online experiment 1 were very unclear (very blurred and very transparent) while the ones of the online experiment 2 were slightly unclear (slightly blurred and slightly transparent). Nevertheless, the video refresh rate was the same for the online experiment 1 and for the online experiment 2.

One more difference was that the dynamic and interactive videos were not replayed in these online studies if no action classification response was recorded to further challenge the accuracy of the participants in the action recognition task, whereas they were replayed in the pilot study up to 30 seconds. The videos are dynamic as the viewpoint can change within the trials and they are also interactive because the changes of the views within the trials depend on the observers' view movements.

In addition, the number of view positions was reduced to 24 (8 angles θ x 3 angles ϕ) in the online studies while they were 60 (12 angles θ x 5 angles ϕ) in the pilot study. This was done to have fewer conditions of starting views and ultimately to reduce the number of trials per participant and make the test shorter for the participants.

Moreover, the test of the online studies was a 4-alternative forced choice task whereas the test of the pilot study was a 6-alternative forced choice task. In the online studies, the participants had to indicate the classes of the actions from four possible actions classes. The four action classes were not the same for all participants, since the four action classes of each participant were randomly selected from a pool of seven possible action classes. I used only four action classes to reduce the number of trials even more. However, in the pilot study, the participants had to classify the actions with one of the six possible action classes. These six possible action classes were the same for all participants of the pilot study.

Another relevant difference was that the participants of the two online studies had one mandatory view movement at the first frame or (timepoint) of each trial and some optional view movements at all remaining frames of each trial, from the second frame to the last one. The first frame of a trail was shown until the participant made the first mandatory view movement. The video was locked at the first frame and no action classification was recorded until the participants selected a mandatory view movement. The first locked frames of the trials in the online experiment 1 were the first frames of the actual videos of the actions from the starting viewpoints, while the first locked frames in the online experiment 2 were the images showing the T-poses of the actors from the starting viewpoints. The starting viewpoints were the views that the trials started with before any view movements. I locked the first frame of each trial to help the participants understand the viewpoint from where they were going to watch the action in any trial before even seeing the action. However, in the pilot experiment of chapter 3, there were three mandatory movements and zero optional view movements, and the dynamic videos were playing normally since their first frames without locking any of them.

The goal of displaying unclear images was to disrupt action recognition. However, a potential side effect of displaying unclear images was that it could also disrupt the self-viewpoint recognition of the participants and the efficient viewpoint selection of the SCM participants. Therefore, I locked the first frame of each trial. In this way, the participants would easily detect the viewpoint from where they were going to watch the action in any trial before even seeing the action. However, by locking the first frame of the actual video until the first view movement like in the online experiment 1, the participants could be exposed to some information about the action for an uncontrolled amount of time. Thus, I locked the T-pose in the online experiment 2 to only display the viewpoint to the participants for an uncontrolled amount of time while revealing no information about the action.

6.2 Experiment 1

6.2.1 Methods

6.2.1.1 Participants

I hired 62 participants on Prolific (www.prolific.co). Their age average was 24.74 (*SD* = 4.42) years and ranged from 18 to 33. 35 participants were male, 26 were females, whereas 1 reported their gender as “Other”. I did not exclude any participant because all participants classified the actions well in terms of accuracy and RT. I randomly assigned 32 participants to the SCM group and 30 to the RM group. I rewarded all participants with money (4.00 £).

6.2.1.2 Materials

I displayed the videos of my dataset MVVHA to assess participants’ action recognition. Specifically, I chose the 2 small versions o.3.6 and ob.3.6 (see chapter 2 for the full descriptions). They both have the same 7 classes of actions, the same 4 actors, the same 27 animation per class, the same 2 mirror conditions, the same 8 angles θ, the same 3 angles ϕ, the same 10 frames per video at 5Hz and the same frame size of 224 x 224 x 3 pixels. Therefore, both o.3.6 and ob.3.6 contains the same 1512 MVVs or the same 36,288 SVVs (1512 MVVs x 24 views). The only difference between these 2 versions is the fact that o.3.6 has clear images (not blurred) whilst, in ob.3.6, all images are blurred. I used o.3.6 for the first 3 familiarizations of SCM procedure and in the first 4 familiarizations of the RM procedure. Instead, I used ob.3.6 for the last 2 familiarizations and for the real test in both the SCM and RM conditions. In this way, the participants started familiarizing with the task with clear images, then they familiarized with blurred images and finally they did the real test with blurred images. See the procedure for the full details of the familiarizations and of the real test.

Each dataset version was stored in a different GitHub repository. Then, to make the SCM and RM tests online, I build 2 Qualtrics (www.qualtrics.com) surveys: one of the SCM test and one for RM test. The two surveys were identical, except for the test question. This test question in the SCM survey had an html frame element that hosted the SCM website that I build with GitHub Pages (pages.github.com). The SCM website runs the SCM test with a html script which has html and JavaScript code. You can visit the SCM website at https://ccalafiore.github.io/demos/demo_2. Likewise, the test question in the RM survey framed the RM website that was also build with GitHub Pages (pages.github.com) which runs the RM test. The RM website is at https://ccalafiore.github.io/demos/demo_3. In both SCM and RM html scripts, I used the JavaScript library jsPsych (www.jspsych.org).

6.2.1.3 Design

This was a mixed design. There were three independent variables: view movement type; the starting angle θ; the starting angle ϕ. The view movement type was manipulated between subjects whilst the starting angle θ and the starting angle ϕ were manipulated within subjects. The view movement type was the control of the view movements which can either be SCM or RM. In SCM group, participants had total control in selecting the next viewpoint, while, in the RM group, the viewpoint changed randomly. The starting views of some action classification trials were the views where the participants started watching the action from in these trials. In this experiment, they started watching the first frame of the video until they make the first view movement. After the first movement, the video started playing from the second frame and the participants could still change the view at any frame. Therefore, the starting views were the ones the participants looked at the actions at the first video frames of the trials. Likewise, the starting angles θ and the starting angles ϕ of some action classification trials were the angles θ and the angles ϕ of the starting views in these trials. There were 24 starting views by combining 8 starting angles θ and 3 starting angles ϕ.

The dependent variables were the action recognition accuracy, the action recognition RT, the number of movements, the percentage frequency of ending angles θ and the percentage of ending angles ϕ. The RTs were timed from the onset of the first frame of the action video before the first movement. The percentage frequencies of ending angles θ and ending angles ϕ of some trials were the frequencies in percentage of the angle θ and the angle ϕ of the ending views of these trials. The ending views of some trials are the ones the participants chose to watch the action from at the end of these trials, just right before their action classifications in these trials.

6.2.1.4 Procedure

The SCM participants did the SCM Qualtrics survey, while the RM participants did the RM Qualtrics survey. Both SCM and RM participants completed their corresponding online survey in about 30 minutes. They were given the link of their corresponding survey. They were instructed and forced to open their link with either Google Chrome, Microsoft Edge or Safari on a laptop or desktop PC.

Both the SCM and RM surveys started with the consent form and, next, they asked the participants their age and their gender. Finally, the SCM survey started the SCM test, whereas the RM survey ran the RM test. The SCM test had 6 phases. The first 5 phases were short familiarizations and the last one was the real test. In the RM, there were 7 phases. The first 6 phases were short familiarizations while the last one was the real test.

In both SCM and RM groups, I randomly selected the 4 classes of actions out of 7 for each participant. The selected classes were the same for familiarizations and real test of the same participant. Additionally, I randomly selected a different actor for each phase. There are 4 actors in o.3.6 and ob.3.6 for 6 phases in the SCM condition or for 7 phases in the RM conditions, so I reshown some actors in different familiarizations. However, the actor that I showed in the real test to a participant was never shown in any familiarization of the same participant. I also randomly chose a different animation per class. There are 27 animations per class which is a lot more than the number of phases, so repetitions of the same animations in different phases were not needed.

The first 3 familiarizations were identical for both the SCM and RM conditions. In the first 3 familiarizations of both SCM and RM conditions and in the fourth familiarization of the RM group, the images were clear (not blurred and not transparent). In the first familiarization, participants had the chance to familiarize with movements of the viewpoints. The trial stared with a fixation cross in the middle of the screen for 1 second plus a 0.5-second jitter. The 0.5-second jitter is an extra random time interval between 0 and 0.5 seconds. Then, a video started with a refresh rate of 5 Hz from a random viewpoint with a random angle θ and a random angle ϕ. From that moment on, the participants were asked to move their own view to the opposed side of the actor. They could move the view by pressing the 4 arrows of their keyboard and confirm with the key space once they reached the opposed view. The video was replayed until they confirmed with space. When they confirmed, the next trials started. The following instructions were given to the participants for the first familiarization.

“Phase Familiarization_0

Here, you will learn how to move the view where you look at the action from by pressing the arrows of your keyboard. You will see an actor doing an action from a random side. Your task is to move your viewpoint to the opposite side of the actor, by pressing the arrows. Once reached, please, press space to continue.”

The second familiarization was the same as the first one. There was only one difference. Here the participants were asked to move to a clear view. A clear view is the view where they believe they can see the action more clearly. Here are the instructions of the second familiarization.

“Phase Familiarization_1

Here, you will still see an actor doing an action. But now, you are asked to move your viewpoint to a clearer one. The clear viewpoints are the ones where you think the action is more clear. You can press any arrow to move to a clear view. Once reached, please, press space to continue.”

The third familiarization was an active action recognition task with clear (not blurred and not transparent) images. In each trial of this familiarization, after showing the fixation cross for 1 seconds plus a jitter of 0.5 seconds, I displayed only the first frame of the 10-frame video from a random viewpoint until the make the first move of their own view. Once the participant moved for the first time, each of the other 9 frames was shown one at the time for 200 milliseconds, reproducing the actual 5-Hz refresh rate of the video. After the first movement, they could move one time for each timepoint (or frame), if they wanted. The trial ended when they classified the action in the video. They could not classify before the first movement. After they moved the first time, they could classify any time before or after the end of the video. They could classify by pressing the keys “1”, “2”, “3”, “4”. The order of the classes was randomised for each participant, but it was the same for all phases of the same participant. In this familiarization, the video of a trial was never replayed, even if the participant had not classified the action at the end of the video. If they had not classified at the end of the video, the text “Which Action?” was shown afterwards instead of the video until action classification. After classification, a feedback screen was displayed to the participant for 1 seconds. The feedback screen either said “Correct!” in green if the classification was correct or “Incorrect!” in red if the classification was wrong. Note that the first movement on the first frame was mandatory, and any further movement on any other frames was optional. In other words, in each trial, the participant could do 1 movement at minimum (during the presentation of the first frame) or 10 movements at maximum (one for each frame). The instructions of the third familiarization were:

“Phase Familiarization_2

In this phase, you are asked to move to a clearer viewpoint and classify the action as quickly and accurate as possible. Classify the action by pressing 1 for Pointing, 2 for Dancing, 3 for Standing Up, 4 for Discussing.

If you did not recognise any action in the video, please, have your best guess as quickly as possible. Just after your classification, you will get a feedback saying "Correct" if your classification was correct or "Incorrect" if your classification was incorrect.”

The fourth SCM familiarization was like the third one with only one difference. The participants saw the videos of ob.3.6 which were blurred. I did not apply any transparency yet on the images in this familiarization. So, the images were blurred and not transparent. They may have seen a different actor and a different animation per class. The action classes and the order of these classes was the same as in all other phases. Everything else was the same as the third familiarization. Here, there was also the feedback screen after classification. The instructions for the fourth SCM familiarization were as follows:

“Phase Familiarization_3

This phase is like the previous one. The only difference is that the images are blurred.”

The fifth SCM familiarization was the same as the fourth SCM familiarization. However, here the images were blurred and transparent. The alpha of the images was 0.1 which mean the blurred images were 10% visible and 90% transparent. The actor and the animation in each class may have been different in this familiarization as the ones in the fourth familiarization. The fifth SCM familiarization was anticipated by the following instructions.

“Phase Familiarization_4

This phase is like the previous one. However, now the images are blurred and transparent.”

The trial in the sixth SCM phase which is the real SCM test was like the ones of the fifth familiarization. The images were blurred and transparent. The actor and the animation per class in the real test were chosen randomly and they were never shown in any familiarization. The instructions of the SCM real test were:

“Phase Real_Experiment

This phase is the Real Experiment. Here, the images are also blurred and transparent. However, you will not get a feedback at the end of every trial.”

The fourth RM familiarization was like the third one, except for the type of movement. The participants saw the clear images of o.3.6 which were not blurred and not transparent like in the third familiarization. However, while the participants had full control of the direction of the view movement in the third familiarization, they had no control of the movement direction in the fourth one because it was random. They still had to press at least one of the 4 arrows of the keyboard to move, but the view actually moved to a random direction (to a random neighbouring view). The first move at the first frame was mandatory here as well, even if it was to a random direction, and any other random move in each of the other 9 frames was optional. The instructions below were shown before the fourth RM familiarization.

“Phase Familiarization_3

This phase is like the previous (third) one. The only difference is the random movements. So, when you move the view, it will move towards a random direction. Your task is still to explore and classify the action as quickly and accurate as possible even if your view moves randomly.”

The fifth RM familiarization was the same as the fourth RM familiarization with the blurred and non-transparent images. The instructions of the fifth RM familiarization are quoted below.

“Phase Familiarization_4

This phase is like the previous one with random movements. The only difference is that the images are blurred.”

The sixth RM familiarization and the RM real test were same as the fifth with blurred and transparent images. However, in the RM real test there were no feedback at the of every trial. The instructions of the sixth RM familiarization are below.

“Phase Familiarization_5

This phase is like the previous one with random movements and with blurred images. However, now, the images are also transparent.”

Instead, the instructions of the RM real test were:

“Phase Real_Experiment

This phase is the Real Experiment. This is like the previous phase with random movements and with blurred and transparent images. However, now, you will not get a feedback at the end of every trial.”

In any SCM or RM familiarization, there were 8 trials, using all possible combinations of the 4 selected action classes, the selected actor, the selected animation per class, the mirrors (4 x 1 x 1 x 2 = 8). Each trial started from a random viewpoint (with a random angle θ and a random angle ϕ). In the real test of both SCM and RM groups, there were 192 trials by combining the 4 selected classes, the selected actor, the chosen animation per class, the 2 mirrors, the 8 possible starting angles θ and the 3 possible starting angles ϕ (4 x 1 x 1 x 2 x 8 x 3 = 192). The starting angle θ and the starting angle ϕ of a trial were the angle θ and the angle ϕ of the viewpoint the video started from. Both the SCM and RM real tests were split in 6 blocks of 32 trials and participants could have a short break between blocks.

You can play the SCM test at https://ccalafiore.github.io/demos/demo_2 and the RM test at https://ccalafiore.github.io/demos/demo_3.

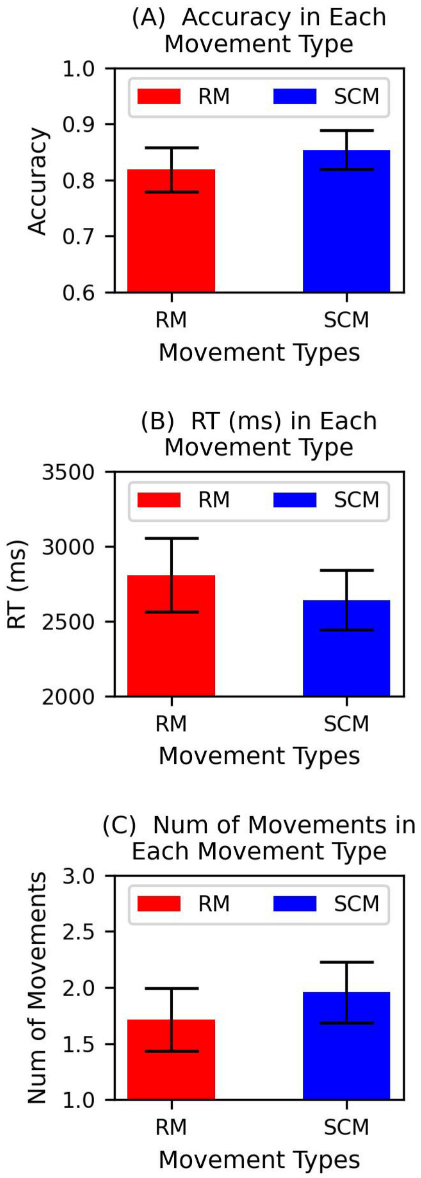


Figure 6.1. The action recognition accuracies (A), the action recognition RTs (B) and the numbers of within-trial view movements (C) in the RM and SCM groups of experiment 1.

6.2.2 Results

6.2.2.1 Effect of the View Motion Type on the Accuracy, the RT and the Number of View Movements

This experiment mainly aimed to unveil whether the action recognition performance of human observers is higher in the SCM condition than in the RM condition in case of very unclear images. If that is the case, the action recognition is expected to be more accurate and shorter in the SCM condition than in the RM condition.

An independent-sample t-test was conducted to compare accuracy in RM and SCM conditions. There was not a significant difference in accuracy for RM (*M* = .819, *SD* = .105) and SCM (*M* = .854, *SD* = .096) conditions; *t*(60) = −1.376, *p* = .174. This result showed participants in the SCM group were not more accurate than the participants in the RM group in the action recognition task. The chart (A) of Figure 6.1 shows the action recognition accuracies of the RM and SCM groups.

An independent-sample t-test was run to compare RT in RM and SCM conditions. There was not a significant difference in RT for RM (*M* = 2,808 ms, *SD* = 660 ms) and SCM (*M* = 2,642 ms, *SD* = 548 ms) conditions; *t*(60) = 1.084, *p* = .283. This suggests SCM participants are not faster than the RM participants in recognizing actions. The RTs of the RM and the SCM groups are displayed in the chart (B) of the Figure 6.1.

I run an independent-sample t-test to compare the numbers of view movements in RM and SCM conditions. There was not a significant difference in number of view movements between the RM (*M* = 1.715, *SD* = .743) and the SCM (*M* = 1.958, *SD* = .75) conditions; *t*(60) = −1.282, *p* = .205. The SCM group moved their view as many times as the RM group. The bar chart (C) in Figure 6.1 pictures the numbers of view movements which were made on average in each trial by the RM and the SCM participants.

6.2.2.2 Effect of the Completed Blocks of Trials on the Accuracy, the RT and the Number of View Movements

The action classification performance of the participants of both RM and SCM groups was expected to improve throughout the real test. Their action classification should have become more accurate and quicker at each block of trials. It is also plausible to reason that the number of within-trial view movements in both groups would decline at each block of trials, because they should have learnt more about how to efficiently reach the target viewpoints with fewer view movements. This argumentation is more appropriate for the

SCM group than the RM group. Therefore, the movement type and the blocks of trials (experience) may interact in determining the number of view movements, since the number of within-trial view movements in the SCM group would decline more than in the RM group at each block of trials.

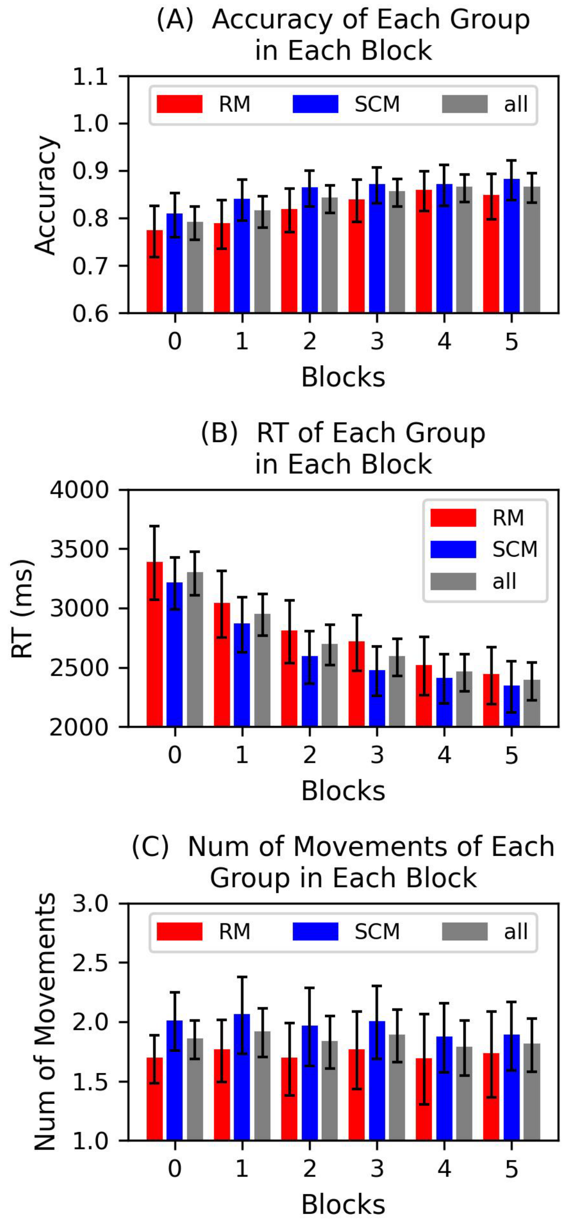


Figure 6.2. The action recognition accuracies (A), the action recognition RTs (B) and the numbers of within-trial view movements (C) in the 6 blocks of 32 trials of experiment 1.

I ran three two-way ANOVA tests on the accuracies, RTs and the view movements of the participants in each of the six blocks of 32 trials to examine the effect of the completed blocks (learning effect) and the interaction effect of the completed blocks and the movement type on the action recognition accuracy, RT and the number of movements. Mauchly's tests showed that the assumptions of sphericity for the accuracies, $\chi^2(14) = 45.865$, $p < .001$, for the RTs, $\chi^2(14) = 92.697$, $p < .001$, and for the numbers of viewpoint movements, $\chi^2(14) = 136.576$, $p < .001$, in the different block conditions had been violated.

Hence, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .76$ for the accuracy; $\varepsilon = .554$ for the RT; $\varepsilon = .479$ for the number of movements).

There was a significant main effect of the block on the action classification accuracy, *F*(3.798, 227.855) = 9.484, *p* < .001, and on the action classification RT, *F*(2.768, 166.103) = 96.798, *p* < .001. Figure 6.2(A) shows that the accuracy in both groups improved almost at each block of trials, whilst Figure 6.2(B) shows that the RT in both groups was shorter and shorter at each block. These results show a considerable learning effect since the participants in both groups processed more accurate and faster action recognitions over the whole test as they get more experience. On the other hand, there was no significant main effect of the block on the amount of view movements, *F*(2.397, 143.81) = .871, *p* = .437. This shows that the number of movements did not vary by the amount of completed blocks of trials. Figure 6.2(C) shows that the numbers of viewpoint movements of all participants of both groups remained stable in the range from 1.776 to 1.908 across the blocks of trials.

Finally, there were no significant interaction effects of the block and the movement type on the action recognition accuracy, *F*(5, 300) = .502, *p* = .775, on the action recognition RT, *F*(5, 300) = .671, *p* = .646, and on the number of movements, *F*(5, 300) = .388, *p* = .857. These results suggest that the changes in accuracy, RT and number of movements across the blocks of trials were similar in both the RM and SCM groups. In short, the learning effect was the same in both RM and SCM participants.

6.2.2.3 Action Recognition Accuracies, Action Recognition RTs and the Numbers of View Movements in Each Action Class

Some action classes, such as dancing, sitting down (standing-to-sitting), standing up (sitting-to-standing) and falling down, can be easily recognised from any viewpoint, while other action classes, like discussing, pointing and waving, can only be detected from a limited range of viewpoints. Therefore, all RM and SCM participants should have recognised dancing, sitting down, standing up and falling down more accurately and more quickly than discussing, pointing and waving. Additionally, both RM and SCM participants should have moved fewer times to classify dancing, sitting down, standing up and falling down than discussing, pointing and waving.

I only computed some descriptive statistics on the action recognition accuracies, the action recognition RTs and the number of view moves of the participants in each of the 7 action classes. I did not calculate any inferential statistics because the study was not designed for it and each participant was only tested on 4 random action classes out of 7.

The recognitions of sitting down (*M* = .937, *SD* = .050), standing up (*M* = .934, *SD* = .062) and falling down (*M* = .942, *SD* = .073) were more accurate than the recognitions of dancing (*M* = .856, *SD* = .193) and waving (*M* = .862, *SD* = .116) which were more accurate than the recognitions of discussing (*M* = .675, *SD* = .234) and pointing (*M* = .668, *SD* = .212).

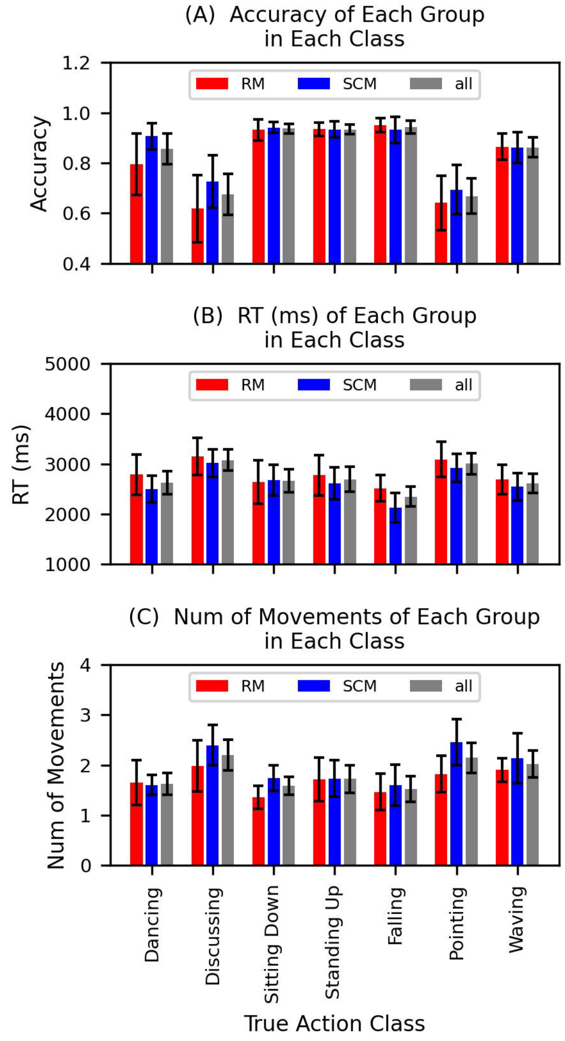
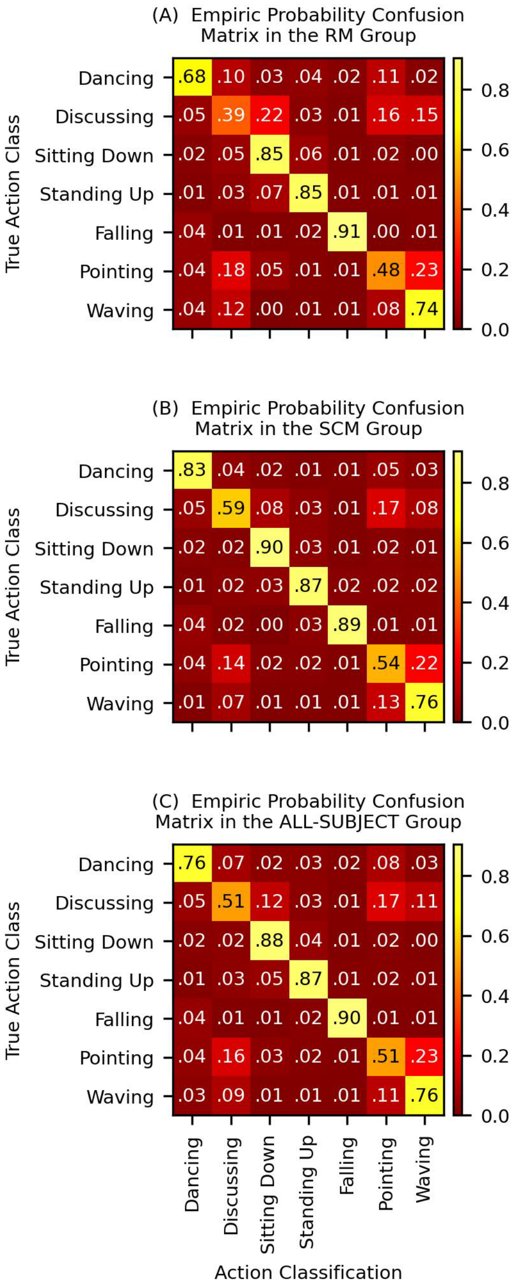


Figure 6.3. The action recognition accuracies (A), the action recognition RTs (B) and the numbers of within-trial view movements (C) in each of the 7 different action classes of the videos. These were the scores of experiment 1.

Figure 6.3(A) highlights the action classification accuracy for every action class.

The classifications of falling down (*M* = 2,352 ms, *SD* = 552 ms) were the quickest ones. These were followed by the classifications of dancing (*M* = 2,628 ms, *SD* = 710 ms), sitting down (*M* = 2,664 ms, *SD* = 600 ms), standing up (*M* = 2,695 ms, *SD* = 772 ms) and waving (*M* = 2,618 ms, *SD* = 578 ms). The classifications of discussing (*M* = 3,079 ms, *SD* = 612 ms) and pointing (*M* = 3,005 ms, *SD* = 643 ms) were the longest. Figure 6.3(B) displays the action classification RTs for all action classes.

The participants moved their viewpoints fewer times when they recognised dancing (*M* = 1.63, *SD* = .67) sitting down (*M* = 1.59, *SD* = .47), standing up (*M* = 1.72, *SD* = .83) and falling down (*M* = 1.53, *SD* = .73) than when they classified discussing (*M* = 2.20, *SD* = .89) pointing (*M* = 2.15, *SD* = .90) waving (*M* = 2.02, *SD* = .82). The numbers of view movements for all action classes are in Figure 6.3(C)



(A) Empiric Probability Confusion Matrix in the RM Group

	Dancing	Discussing	Sitting Down	Standing Up	Falling	Pointing	Waving
Dancing	.68	.10	.03	.04	.02	.11	.02
Discussing	.05	.39	.22	.03	.01	.16	.15
Sitting Down	.02	.05	.85	.06	.01	.02	.00
Standing Up	.01	.03	.07	.85	.01	.01	.01
Falling	.04	.01	.01	.02	.91	.00	.01
Pointing	.04	.18	.05	.01	.01	.48	.23
Waving	.04	.12	.00	.01	.01	.08	.74

(B) Empiric Probability Confusion Matrix in the SCM Group

	Dancing	Discussing	Sitting Down	Standing Up	Falling	Pointing	Waving
Dancing	.83	.04	.02	.01	.01	.05	.03
Discussing	.05	.59	.08	.03	.01	.17	.08
Sitting Down	.02	.02	.90	.03	.01	.02	.01
Standing Up	.01	.02	.03	.87	.02	.02	.02
Falling	.04	.02	.00	.03	.89	.01	.01
Pointing	.04	.14	.02	.02	.01	.54	.22
Waving	.01	.07	.01	.01	.01	.13	.76

(C) Empiric Probability Confusion Matrix in the ALL-SUBJECT Group

	Dancing	Discussing	Sitting Down	Standing Up	Falling	Pointing	Waving
Dancing	.76	.07	.02	.03	.02	.08	.03
Discussing	.05	.51	.12	.03	.01	.17	.11
Sitting Down	.02	.02	.88	.04	.01	.02	.00
Standing Up	.01	.03	.05	.87	.01	.02	.01
Falling	.04	.01	.01	.02	.90	.01	.01
Pointing	.04	.16	.03	.02	.01	.51	.23
Waving	.03	.09	.01	.01	.01	.11	.76

Figure 6.4. The confusion matrices of the RM (A), SCM (B) and all participants (C) in experiment 1.

The confusion matrix in Figure 6.4(C) shows that the participants tended to confuse pointing and waving at the highest rate. They confused discussing and pointing at the second highest rate. The third, fourth and fifth largest confusing rates of the participants were for discussing and waving, for dancing and pointing, and for dancing and discussing, respectively.

6.2.2.4 Evidence of the Efficient and Inefficient Views from the Starting Views

This study also aimed to show again the efficient and inefficient views for the action recognition of human observers, by using the starting views in both RM and SCM conditions. I assumed that the action classification of the participants in both RM and SCM conditions would be more accurate and faster when their starting views were efficient than when their starting views were inefficient. Similarly, I also assumed that the participants in both RM and SCM would move less times when their starting views were efficient than when their starting views were inefficient.

6.2.2.4.1 Effect of the Starting View on the Accuracy, the RT and the Number of View Movements

I conducted three one-way ANOVAs on the accuracies, RTs and the numbers of view movements of all RM and SCM participants in all 24 starting views to investigate the effects of the starting view on the action recognition accuracy, RT and the quantity of view movements. Mauchly's tests showed that the assumption of sphericity for the RTs, $\chi^2(275) = 333.568$, $p = .013$, and the view movements, $\chi^2(275) = 449.62$, $p < .001$, had been violated. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .661$ for the RT; $\varepsilon = .601$ for the view movements). I did not follow up on any of the significant effects with t-tests because there were too many (276) pairs of 24 conditions with different starting views for each of the three dependent variables.

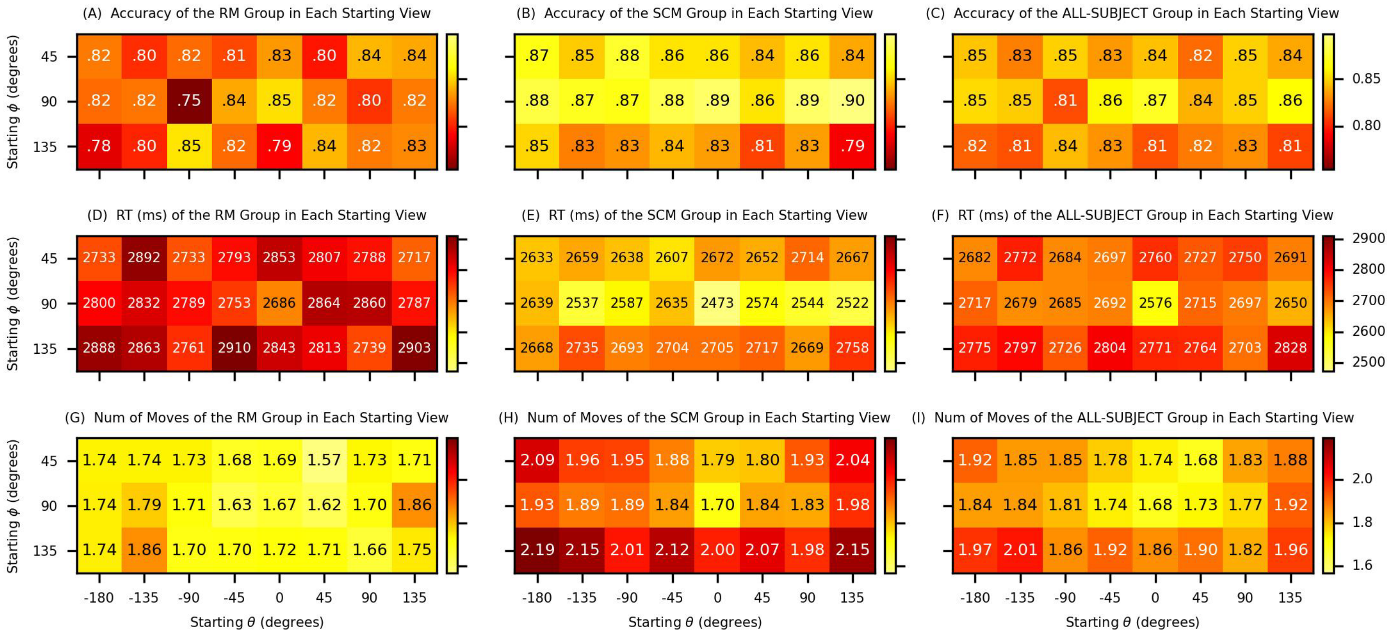


Figure 6.5. The accuracies of the RM (A), SCM (B) and all (C) participants in each starting view. The RTs of the RM (D), SCM (E) and all (F) participants in every starting view. The quantity of the view movements of the RM (D), SCM (E) and all (F) participants in every starting view. All these results come from experiment 1.

There was a marginal effect of the starting view on the accuracy, *F*(23, 1403) = 1.409, *p* = .094. Some views are slightly more efficient than others given that the action recognition accuracy was marginally affected by the starting view. Figure 6.5(C) shows the accuracies of all participants of both groups given each starting view. There was a significant effect of the starting view on the RT, *F*(15.207, 927.648) = 2.427, *p* = .002. Because the action classification speed was affected by the starting view, then we can reason the distinction between the efficient and inefficient views. Figure 6.5(F) illustrates the RTs of all participants in each starting view.

There was a significant effect of the starting view on the number of view movements, *F*(13.825, 843.302) = 4.353, *p* < .001. It seems like the participants moved their viewpoints away from the inefficient views to the efficient ones since the numbers of view movements tended to be more when the starting views were inefficient than when they were efficient. The heatmap in Figure 6.5(I) highlights the numbers of the within-trial view movements in each of the 24 starting views.

6.2.2.4.2 Effect of the Starting Angle ɸ on the Accuracy, the RT and the Number of View Movements

Let us now continue by looking at whether the starting angle ɸ effected the accuracy, the RT and the number of movements of both RM and SCM participants.

A one-way within-subject ANOVA was conducted to assess the effect of the starting angle ɸ on accuracies in the top ($ɸ_0$=45°), middle ($ɸ_1$=90°) and bottom ($ɸ_2$=135°) starting views. There was a significant effect of the starting angle ɸ on accuracy, *F*(2, 122) = 5.282, *p* = .006. I followed up on the effect with multiple paired t-tests with the Bonferroni correction between all conditions of starting angles ɸ. Action Recognition with bottom starting angle ɸ (*M* = .823, *SD* = .107) was significantly less accurate than with the middle angle ɸ (*M* = .85, *SD* = .112), *t*(61) = 3.032, *p* = .011, and it was not significantly less accurate than with the top starting angle ɸ (*M* = .838, *SD* = .104), *t*(61) = 1.97, *p* = .160. On the other hand, action recognition accuracy with top starting angle ɸ and with the middle starting angle ɸ was not significantly different, *t*(61) = −1.399, *p* = .501. These results on the accuracy of the starting angles ɸ suggest that the middle angles ɸ is efficient for action recognition while the bottom angle ɸ is inefficient. That is because action recognition was more accurate when the stimulation of the action started from the middle views than when it started from the bottom ones. The accuracies of all groups in all starting view angles ɸ are shown in the graph (A) of Figure 6.6.

A one-way within-subject ANOVA was conducted to compare the effect of the starting angle ɸ on RT in top ($ɸ_0$=45°), middle ($ɸ_1$=90°) and bottom ($ɸ_2$=135°) starting views. There was a significant effect of angle ɸ on RT, *F*(2, 122) = 12.86, *p* < .001. Therefore, action recognition was faster in some starting angles ɸ than others. I followed up on this effect by running multiple paired t-tests with the Bonferroni correction to compare the RT in all 3 starting angles ϕ. The bottom views ϕ_2=135° (*M* = 2,771 ms, *SD* = 624 ms) were significantly slower than the top views ϕ_0=45° (*M* = 2,720 ms, *SD* = 606 ms), *t*(61) = −2.991, *p* = .012, and middle views ϕ_1=90° (*M* = 2,676 ms, *SD* = 605 ms), *t*(61) = −4.76, *p* < .001. However, the middle views ϕ_1=90° were not significantly faster than the top views ϕ_0=45°, *t*(61) = 2.293, *p* = .076. These results suggest that the efficient angle ϕ is the middle one because participants were faster when they started the trials with the middle views. On the other hand, the inefficient angle ϕ is the bottom one as action recognition was slower when participants started watching from bottom views. The bar chart (B) of Figure 6.6 highlights the RTs of all participants in every starting view angle ϕ.

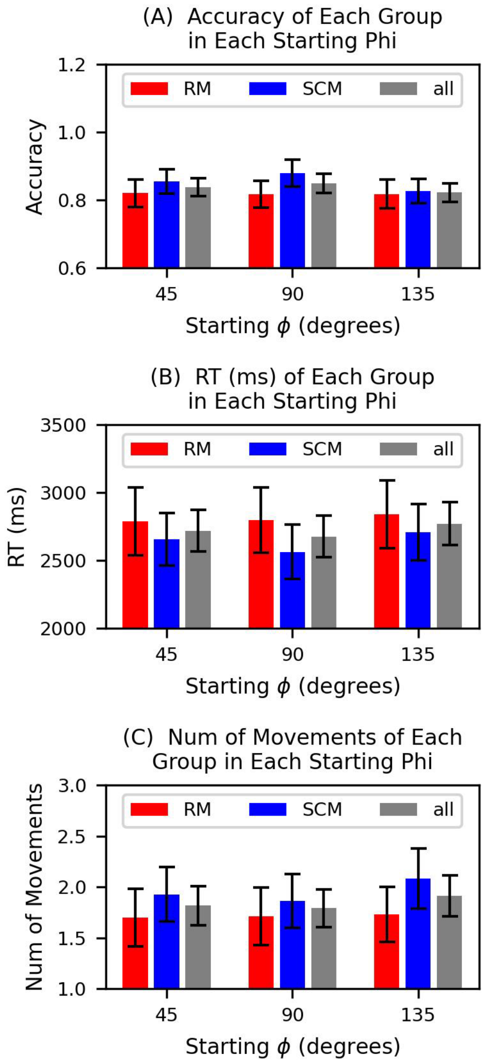


Figure 6.6. The accuracies (A), the RTs (B) and the quantity of the view movements (C) of the RM, SCM and all participants in each starting view angle ϕ. These results come from experiment 1.

Furthermore, a one-way within-subject ANOVA was conducted to compare the effect of the starting angle ϕ on the number of movements in top (ϕ_0=45°), middle (ϕ_1=90°) and bottom (ϕ_2=135°) starting views. There was a significant effect of angle ϕ on the number of movements, *F*(2, 122) = 12.235, *p* < .001. I run multiple pairwise comparisons by paired t-tests with the Bonferroni correction to further investigate the latter effect on the number movements. Participants moved their view significantly more times if they started watching from the bottom views ϕ_2=135° (*M* = 1.913, *SD* = .788) than if they started watching from the top views ϕ_0=45° (*M* = 1.817, *SD* = .751), *t*(61) = −3.414, *p* = .003, and the middle views ϕ_1=90° (*M* = 1.791, *SD* = .739), *t*(59) = −4.964, *p* < .001. The participant with the top starting views approximately moved as many time as the middle starting view, *t*(61) = 1.018, *p* = .938. These results reveal that efficient views were the top and middle ones, while the inefficient views were the bottom views. That is because the participants moved less times when they started the trials from top and middle views, whereas they made more movement when they started the trial from the bottom views. The numbers of view movements of the RM, SCM and all participants in each starting view angle ϕ are displayed in the plot (C) of Figure 6.6.

6.2.2.4.3 Effect of the Starting Angle θ on the Accuracy, the RT and the Number of View Movements

To highlight the efficient and inefficient angles θ, I also investigated the effect of angle θ on the accuracy, RT and number of movements in all starting angles θ.

A one-way within-subject ANOVA was conducted to compare the effect of the starting angle θ on accuracy in all eight conditions. There was not a significant effect of the starting angle θ on accuracy, *F*(7, 427) = .469, *p* = .857. The accuracy did not vary, by varying the starting angle θ and did not highlight the efficient and inefficient angles θ. Figure 6.7(A) shows the accuracies in each starting view angle θ for the RM and SCM groups and for both.

A one-way within-subject ANOVA was conducted to compare the effect of the starting angle θ on RT in all eight conditions. There was not a significant effect of the starting angle θ on RT, *F*(7, 427) = .792, *p* = .594. Therefore, the starting angle θ did not have an effect even on the RT. The RTs in each starting view angle θ for the RM, SCM and both groups can be seen in the bar chart (B) of Figure 6.7.

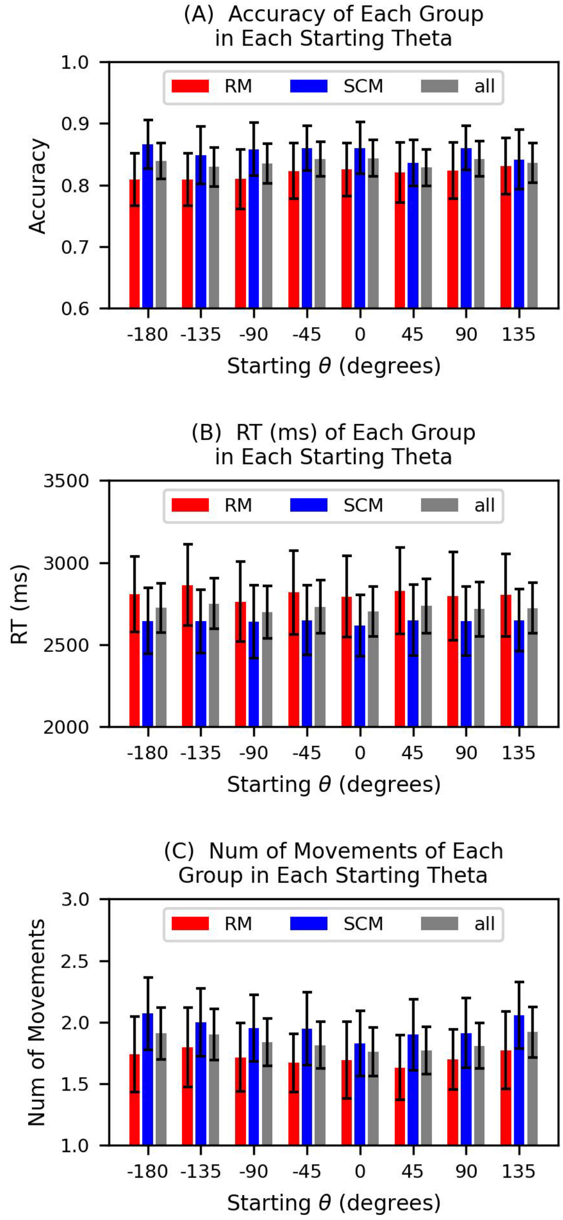


Figure 6.7. The accuracies (A), the RTs (B) and the quantity of the view movements (C) of the RM, SCM and all participants in each starting view angle θ. These were the results of experiment 1.

A one-way within-subject ANOVA was conducted to compare the effect of the starting angle θ on the number of movements in all eight conditions of the angle θ. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(27) = 88.266$, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .696$). There was a significant main effect of the starting angle θ on the number of movements, $F(4.869, 296.985) = 5.405$, $p < .001$. I further investigated this effect with multiple paired t-tests with the Bonferroni correction. These showed the participants significantly made more movements when they started watching from the back $\theta_0 = -180°$ (*M* = 1.91, *SD* = .829), right-back $\theta_1 = -135°$ (*M* = 1.901, *SD* = .814) and left-back $\theta_7 = 135°$ (*M* = 1.919, *SD* = .803) views than when they began watching from front $\theta_4 = 0°$ (*M* = 1.761, *SD* = .781), left-front $\theta_5 = 45°$ (*M* = 1.77, *SD* = .761) views. The efficient angles θ seemed to be the front views because participants moved less times

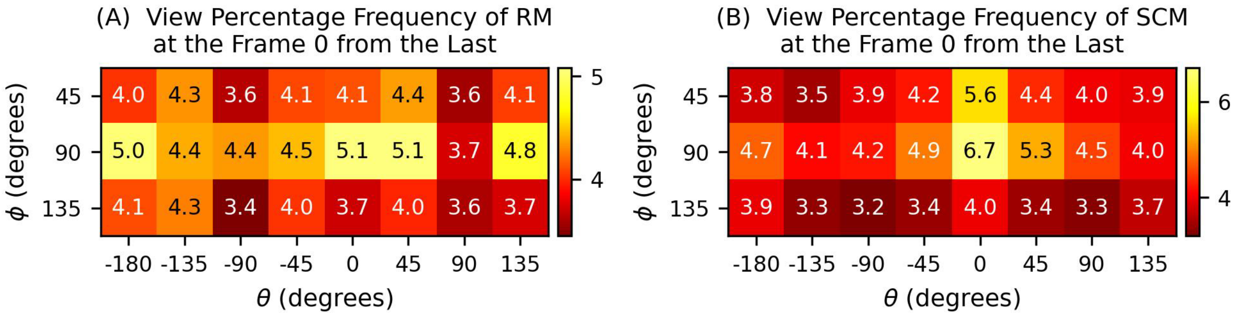


Figure 6.8. The percentage frequencies of the selected ending views in the RM and the SCM groups of experiment 1.

when they started the trial from the front views. That is because they were already in a view with an efficient angle θ and did not need to move a lot. However, the inefficient angles θ were the back ones as they made more view moves when they started watching from the back views. This might be because participants needed more moves on their way from bad views to good views. Figure 6.7(C) pictures the numbers of within-trial view movements of the participants for all starting angles θ.

6.2.2.5 Evidence of Efficient View Selection

One of my objectives of this experiment was to examine whether the view selection of the SCM participants for action recognition was efficient. Thus, I explored here whether the human participants in the SCM condition tended to select the efficient views more often than the inefficient views.

6.2.2.5.1 View Selection

A one-way ANOVA was run on the percentage frequencies of all ending views of the SCM group to assess whether the SCM participants selected some views more often than some other views to watch the actions at the last timepoint of the video stimulation, just before their action classifications. Mauchly’s test showed that the assumption of sphericity had been violated, $\chi^2(275) = 685.669$, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .240). The percentage frequencies significantly varied across the different ending views, $F(5.529, 171.396) = 4.227$, $p < .001$. This result shows that the SCM participants chose some views more often than others to see the action at the last timepoint of the video, just before their action classification. Figure 6.8(B) shows the percentage frequencies of views at the last timepoint of all trials. Front views were preferred over the back and side views, as well as the middle views were selected more than the top and bottom views.

6.2.2.5.2 Angle ϕ Selection

I ran an ANOVA to compare the percentage frequencies of the ending angles ϕ in the SCM condition. The ending view angles ϕ could only be top, middle and bottom. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(2) = 7.536$, $p = .023$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .818). The percentage frequencies of the ending angles ϕ were significantly different, $F(1.636, 50.731) = 5.642$, $p = .009$. I followed up on the effect by running multiple pairwise paired t-tests with the Bonferroni correction. While the SCM participants did not select the middle angle ϕ (M = 38.517, SD = 7.284) significantly more often than the top angle ϕ (M = 33.269, SD = 11.127), $t(31) = -1.959$, $p = .177$, they did select the middle angle significantly more often than the bottom angle ϕ (M = 28.214, SD = 11.143), $t(31) = 3.838$, $p = .002$. Furthermore, the selection of the bottom angle ϕ and the top angle ϕ were not

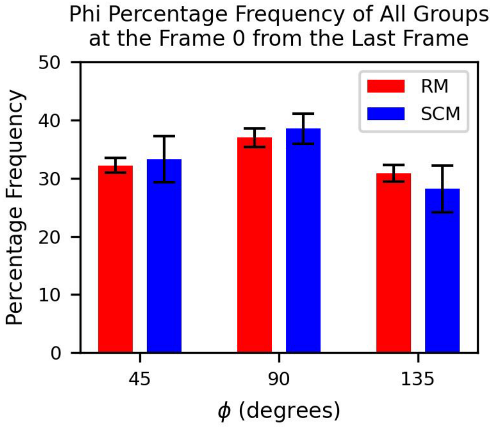


Figure 6.9. The percentage frequencies of the selected ending view angle ϕ in the RM and the SCM groups of experiment 1.

significantly different, *t*(31) = 1.359, *p* = .552. This pattern can be clearly seen in the Figure 6.9. In summary, the SCM participants selected more often the efficient angle ϕ which was the middle one. They also selected less often the inefficient angle ϕ which was the bottom one.

6.2.2.5.3 Angle θ Selection

Next, I looked at whether people in the SCM condition selected more often the efficient angles θ than the inefficient angles θ of the views. Therefore, I conducted an ANOVA to compare the percentage frequencies of all eight ending angles θ in the SCM condition. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(27)$ = 172.301, *p* < .001. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (*ε* = .301). The percentage frequencies of the ending angles θ were significantly different, *F*(2.110, 65.420) = 4.394, *p* = .015. I followed up on the effect with multiple pairwise paired t-tests between the percentage frequencies of the eight possible ending angles θ. The p-values were corrected by the Bonferroni method. These showed that participants significantly selected the front view with θ_4=0° (*M* = 16.276, *SD* = 6.624) and left-front view with θ_5=0° (*M* = 13.034, *SD* = 2.232) more often than the back (*M* = 12.443, *SD* = 5.401) and the right-back (*M* = 10.928, *SD* = 1.940) view angles θ. Figure 6.10 shows the percentage frequencies of the selected ending views for both RM and SCM participants. Thus, the participants in the SCM group choose the efficient angle θ

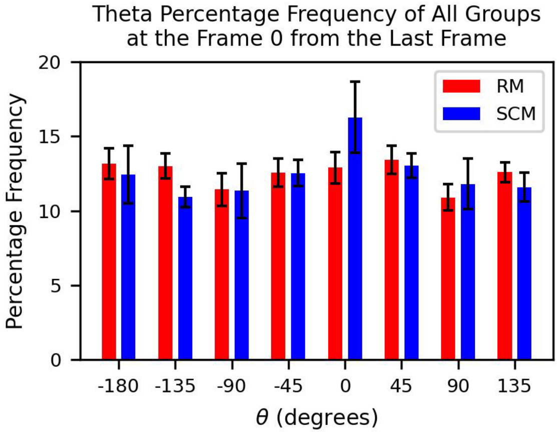


Figure 6.10. The percentage frequencies of the selected ending view angle θ in the RM and the SCM groups of experiment 1.

more frequently than the inefficient angles θ.

6.2.2.5.4 Correlations Between the Selection of Views and the Action Recognition Performances of the same Views

A Pearson correlation coefficient was computed to assess the relationship between the percentage frequency of the selected ending views in the SCM group (Figure 6.8(B)) and the view accuracy of the no-motion (NM) group of study in chapter 4 (Figure 4.2(A)). In the NM study, I excluded the left views and there were 15 views (5 angles θ x 3 angles ϕ), whereas, in this study, there were 24 (8 angles θ x 3 angles ϕ). By assuming that the right and left views were symmetric, I defined the accuracy in left-back views as the accuracy in the right-back views, the accuracy in the left views as the accuracy in the right views, and the accuracy in the left-front views as the accuracy in the right-front ones. In this way, I made 24 datapoints for the correlation: 24 view percentage frequencies and 24 view

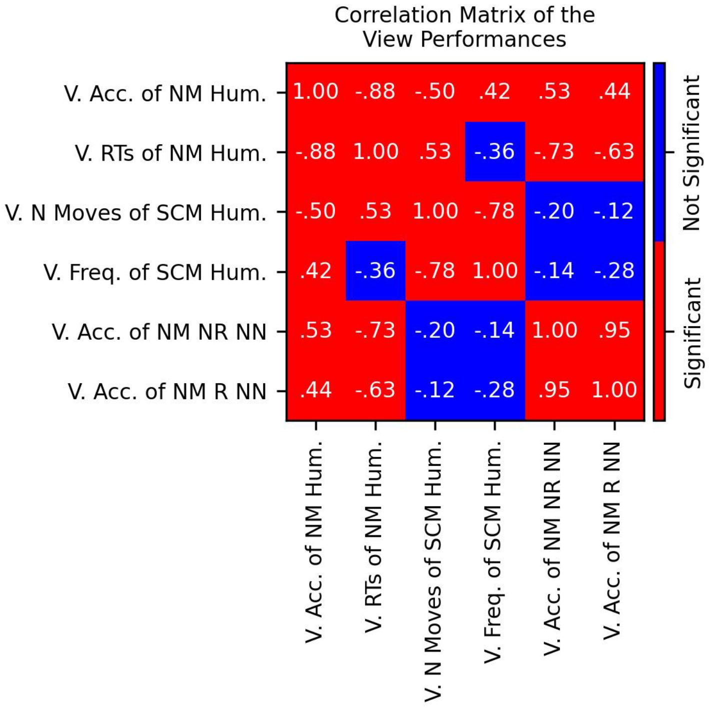


Figure 6.11. The correlation matrix of the following view performances: the view accuracies (V. Acc.) and the view RTs (V. RTs) of the human participants with no view motion (NM Hum.) from the study of the chapter 4; the numbers of view movements in each starting view (V. N Moves) and the percentage frequencies of the ending views (V. Freq.) in the SCM human participants (SCM Hum.) in experiment 1 of this chapter 6; the view accuracies (V. Acc.) of the non-recurrent (NR) and the recurrent (R) models (NN) of the study in chapter 5 with no viewpoint movements (NM).

accuracies. As in the correlation matrix of Figure 6.11, there was a significant positive correlation between the two variables, *r*(22) = .42, *p* = .041. Overall, there was a strong, positive correlation between SCM view selection and the NM view accuracy. This result shows that SCM participants selected the views efficiently as they selected more often the efficient views with higher NM accuracy than the inefficient views with lower NM accuracy.

A Pearson correlation coefficient was computed to assess the relationship between the percentage frequency of the ending views in the SCM group in Figure 6.8(B) and the NM view RTs in Figure 4.2(B). Again, in the NM study, I excluded the left views and there were 15 views (5 angles θ x 3 angles ϕ), whereas, in this study, there were 24 (8 angles θ x 3 angles ϕ). By assuming that the right and left views were symmetric, I defined RTs in left-back, left, left-front views equal to the RTs in the right-back, right, right-front views, respectively. In this way, I made 24 datapoints for the correlation: 24 view percentage frequencies and 24 view RTs. Figure 6.11 also shows that there was a marginal negative correlation between the two variables, *r*(22) = −.36, *p* = .082. This result suggests that SCM participants selected the views efficiently because they selected more often the efficient views with shorter NM RTs than the inefficient views with lower NM RT.

6.3 Experiment 2

6.3.1 Methods

6.3.1.1 Participants

I hired 60 participants on Prolific (www.prolific.co). Their age average was 24.47 (*SD* = 4.23) years and ranged from 18 to 35 years. 31 participants were male, 28 were females, and 1 participant reported their gender is “Other”. I did not exclude any participant because all participants classified the actions well in terms of accuracy and RT. I randomly assigned 30 participants to the SCM group and 30 to the RM group. I rewarded all participants with money (3.50 £).

6.3.1.2 Materials

I displayed the videos of my dataset MVVHA to assess participants' action recognition. For this study, I only used the small versions o.3.6 (chapter 2). The o.3.6 has 7 classes of actions, the same 4 actors, the same 27 animation per class, the same 2 mirror conditions, the same 8 angles θ, the same 3 angles ϕ, the same 10 frames per video at 5Hz and the same frame size of 224 x 224 x 3 pixels. Therefore, o.3.6 contains 1,512 MVVs or 36,288 SVVs (1,512 MVVs x 24 views). In o.3.6, the images are clear which I slightly blurred and made slightly transparent on-the-fly in some familiarizations and in the real test. The o.3.6 also contains the multi-view images of the T-poses. It has 96 T-pose images (4 actors x 8 angles θ x 3 angles ϕ).

The o.3.6 was stored in a GitHub repository. Then, to make the SCM and RM tests online, I build 2 Qualtrics (www.qualtrics.com) surveys: one for the SCM test and one for the RM test. The two surveys were identical, except for the test question. This test question in the SCM survey had an html frame element that hosted the SCM website that I build with GitHub Pages (pages.github.com). The SCM website runs the SCM test with a html script which has html and JavaScript code. You can visit the SCM website at https://ccalafiore.github.io/demos/demo_4. Likewise, the test question in the RM survey framed the RM website that was also build with GitHub Pages and runs the RM test. The RM website is at https://ccalafiore.github.io/demos/demo_5. In both SCM and RM html scripts, I used the JavaScript library jsPsych (www.jspsych.org).

6.3.1.3 Design

This was a mixed design. There were three independent variables: view movement type; the starting angle θ; the starting angle ϕ. The view movement type was manipulated between subjects whilst the starting angle θ and the starting angle ϕ were manipulated within subjects. The view movement type was the control of the view movements which can either be SCM or RM. In SCM group, participants had total control in selecting the next viewpoint, while, in the RM group, the viewpoint changed randomly. The starting angle θ and the starting angle ϕ were the angle θ and the angle ϕ of the view where the participants started watching the action from. In this version of the study, they started watching the T-pose of the actor until they make the first view movement. After the first movement, the video of the action started playing and the participants could still change the view at any frame. Therefore, the starting angle θ and the starting angle ϕ were the parameters of the view they saw the first T-pose frame. There were 8 starting angles θ and 3 starting angles ϕ, which made 24 starting views.

The dependent variables were the action recognition accuracy, the action recognition RT, the number of movements, the percentage frequency of ending angles θ and the percentage of ending angles ϕ. The RTs were timed from the onset of the first frame of the action video after the T-pose and the first movement. The percentage frequencies of the ending angles θ and ending angles ϕ of some trials were the frequencies in percentage of the angles θ and the angles ϕ of the starting views of the trials. The starting views of some trials were the views which the participants selected to watch the actions from at the end of the trials, just right before their action classifications in these trials.

6.3.1.4 Procedure

The SCM participants did the SCM Qualtrics survey, while the RM participants did the RM

Qualtrics survey. Both SCM and RM participants completed their corresponding online survey in about 30 minutes. They were given the link of their corresponding survey. They were instructed and forced to open their link with either Google Chrome, Microsoft Edge or Safari on a laptop or desktop PC.

Both the SCM and RM surveys started with the consent form and, next, they asked the participants their age and their gender. Finally, the SCM survey started the SCM test, whereas the RM survey ran the RM test. The SCM test had 6 phases. The first 5 phases were short familiarizations and the last one was the real test. In the RM, there were 7 phases. The first 6 phases were short familiarizations while the last one was the real test.

In both SCM and RM groups, I randomly selected the 4 classes of actions out of 7 for each participant. The selected classes were the same for familiarizations and real test of the same participant. Additionally, I randomly selected a different actor for each phase. There are 4 actors in o.3.6 for 6 phases in the SCM condition or for 7 phases in the RM conditions, so I reshown some actors in different familiarizations. However, the actor that I showed in the real test to a participant was never shown in any familiarization of the same participant. I also randomly chose a different animation per class. There are 27 animations per class which is a lot more than the number of phases, so repetitions of the same animations in different phases were not needed.

The first 3 familiarizations were identical for both the SCM and RM conditions. In the first 3 familiarizations of both SCM and RM conditions and in the fourth familiarization of the RM group, the images were clear (not blurred and not transparent). In the first familiarization, participants had the chance to familiarize with movements of the viewpoints. The trial stared with a fixation cross in the middle of the screen for 1 second plus a 0.5-second jitter. The 0.5-second jitter is an extra random time interval between 0 and 0.5 seconds. Then, the T-pose was displayed with the hips in the middle of the screen (or in the same position of the previous fixation cross). The T-pose was presented from a view with one out of 8 starting angles θ and one out of 3 starting angles ϕ until the participant makes the first mandatory view movement. Following the view movement, a video started with a refresh rate of 5 Hz from view the participant moved to. The participants were asked to move their own view to the opposed side of the actor. They could move the view by pressing the 4 arrows of their keyboard and confirm with the key space once they reached the opposed view. The video was replayed until they confirmed with space. When they confirmed, the next trials started.

The second familiarization was the same as the first one. There was only one difference. Here the participants were asked to move to a clear view. A clear view is the view where they believe they can see the action more clearly.

The third familiarization was an active action recognition task with clear (not blurred and not transparent) images. In each trial of this familiarization, after showing the fixation cross for 1 seconds plus a jitter of 0.5 seconds, I also displayed the T-pose from a random viewpoint until they make the first move of their own view. Once the participant moved for the first time, one of the 10-frame videos was shown one at the time for 200 milliseconds, reproducing the 5-Hz refresh rate. After the first movement, they could move one time for each timepoint (or frame), if they wanted. The trial ended when they classified the action in the video. They could not classify before the first movement. After they moved the first time, they could classify any time before or after the end of the video. They could classify by pressing the keys “1”, “2”, “3”, “4”. The order of the classes was randomised for each participant, but it was the same for all phases of the same participant. In this familiarization, the video of a trial was never replayed, even if the participant had not classified the action at the end of the video. If they had not classified at the end of the video, the text “Which Action?” was shown afterwards instead of the video until action classification. After classification, a feedback screen was displayed to the participant for 1 seconds. The feedback screen either said “Correct!” in green if the classification was correct or “Incorrect!” in red if the classification was wrong. Note that the first movement at the T-pose frame was mandatory, and any further movement at any 10 video frames was optional. In other words, in each trial, the participant could do 1 movement at minimum (at the presentation of the T-pose) or 10 movements at maximum (one for each frame).

The fourth SCM familiarization was like the third one with only one difference. The T-pose was still clear, but the image of the video were slightly blurred images. I did not apply any transparency yet on the images in this familiarization. So, the images were blurred and not transparent. They may have seen a different actor and a different animation per class. The action classes and the order of these classes was the same as in all other phases. Everything else was the same as the third familiarization. Here, there was also the feedback screen after classification.

The fifth SCM familiarization was the same as the fourth SCM familiarization. However, here the images were blurred and transparent. The alpha of the images was 0.1 which means the blurred images were 10% visible and 90% transparent. The actor and the animation in each class may have been different in this familiarization as the ones in the fourth familiarization.

The trial in the sixth SCM phase which is the real SCM test was like the ones of the fifth familiarization. The images were blurred and transparent. The actor and the animation per class in the real test were chosen randomly and they were never shown in any familiarizations.

The fourth RM familiarization was like the third one, except for the type of movement. The participants saw the clear images of o.3.6 which were not blurred and not transparent like in the third familiarization. However, while the participants had full control of the direction of the view movement in the third familiarization, they had no control of the movement direction in the fourth one because it was random. They still had to press at least one of the 4 arrows of the keyboard to move, but the view actually moved to a random direction (or to a random neighbouring view). The first move at the T-pose frame was mandatory here as well, even if it was to a random direction, and any other random move in each of the 10 video frames was optional.

The fifth RM familiarization was the same as the fourth RM familiarization with the blurred and non-transparent images. The sixth RM familiarization and the RM real test were same as the fifth with blurred and transparent images. However, in the RM real test there were no feedback at the of every trial.

The instructions of first and second SCM and RM phases in experiment 2 were identical to ones in the experiment 1. The instructions of all other phases were also the same in experiment 1 and 2, except for the following additional request:

> *“Please, do not always move the view to the same direction to classify the action even faster. You need to move towards a clear view and then classify the action as quickly as possible."*

In experiment 1, I noticed that a few participants moved always to the same direction to speed up their classifications. Therefore, I added this extra request in the instructions to avoid that some participants would have done the same.

There were 8 trials in any SCM or RM familiarization, by combining the 4 selected action classes, the selected actor, the selected animation per class, the 2 mirrors (4 x 1 x 1 x 2 = 8). Each trial started from a random viewpoint (with a random angle θ and a random angle ϕ). In the real test of both SCM and RM groups, there were 192 trials by combining the 4 selected classes, the selected actor, the chosen animation per class, the 2 mirrors, the 8 possible starting angles θ and the 3 possible starting angles ϕ (4 x 1 x 1 x 2 x 8 x 3 = 192). The starting angle θ and the starting angle ϕ of a trial were the angle θ and the angle ϕ of the viewpoint the T-pose was seen from. Both the SCM and RM real tests were split in 6 blocks of 32 trials and participants could have a short break between blocks.

You can run the SCM test at https://ccalafiore.github.io/demos/demo_4 and the RM test at https://ccalafiore.github.io/demos/demo_5.

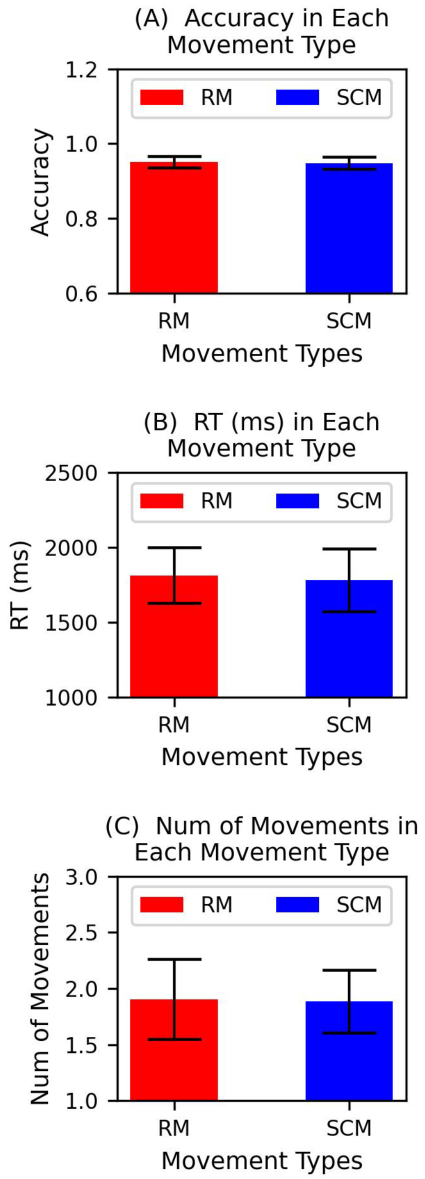


Figure 6.12. The action recognition accuracies (A), the action recognition RTs (B) and the numbers of within-trial view movements (C) in the RM and SCM groups of experiment 2.

6.3.2 Results

6.3.2.1 Effect of the View Motion Type on the Accuracy, the RT and the Number of View Movements

An aim of this experiment was to study whether the action recognition performance of people is higher in the SCM condition than in the RM condition in case of very unclear images. Thus, I specifically looked at whether the action recognition is more accurate and shorter in the SCM condition than in the RM condition.

An independent-sample t-test was conducted to draw a comparison between the accuracies in RM and SCM conditions. There was not a significant difference in accuracy for RM (*M* = .951, *SD* = .042) and SCM (*M* = .948, *SD* = .043) conditions; *t*(58) = .225, *p* = .823. Therefore, participants in the SCM group were not more accurate than the participants in the RM group on the action recognition task.

An independent-sample t-test was run to compare RT in RM and SCM conditions. There was no significant difference in RT for RM (*M* = 1,813 ms, *SD* = 500 ms) and SCM (*M* = 1,781 ms, *SD* = 566 ms) conditions; *t*(58) = .226, *p* = .822. Hence, the SCM participants were not faster than the RM participants in recognizing actions.

I also ran an independent-sample t-test to compare the numbers of view movements in RM and SCM conditions. No significant difference was found in number of movements for RM (*M* = 1.905, *SD* = .959) and SCM (*M* = 1.887, *SD* = .751) conditions; *t*(58) = .077, *p* = .939. The SCM group moved their view as many times as the RM group.

The bar charts (A), (B) and (C) in Figure 6.12 display the accuracies, the RTs and the number of view movements of the RM and the SCM groups of the experiment 1.

6.3.2.2 Effect of the Completed Blocks of Trials on the Accuracy, the RT and the Number of View Movements

Improvement of the action classification performance was expected from the participants of both RM and SCM groups throughout the real test. At each block of trials, the participants should have processed more accurate and faster action classifications as they got more practice with the task. The number of within-trial view movements in both groups should have also declined at each block of trials, as they should have had learnt with more experience how to efficiently reach the target viewpoints with fewer view movements.

Three two-way ANOVA tests were run on the accuracies, RTs and the view movements of the participants in each of the 6 blocks of 32 trials to assess the effect of the completed blocks (learning effect) and the interaction effect of the blocks and the movement type on the action recognition accuracy, RT and the number of viewpoint moves. Mauchly's tests indicated that the assumptions of sphericity for the accuracies, $\chi^2(14) = 43.227$, $p < .001$, for the RTs, $\chi^2(14) = 120.596$, $p < .001$, as well as for the numbers of viewpoint movements, $\chi^2(14) = 140.185$, $p < .001$, in the different block conditions had been violated. Hence, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .798$ for the accuracy; $\varepsilon = .501$ for the RT; $\varepsilon = .497$ for the number of movements).

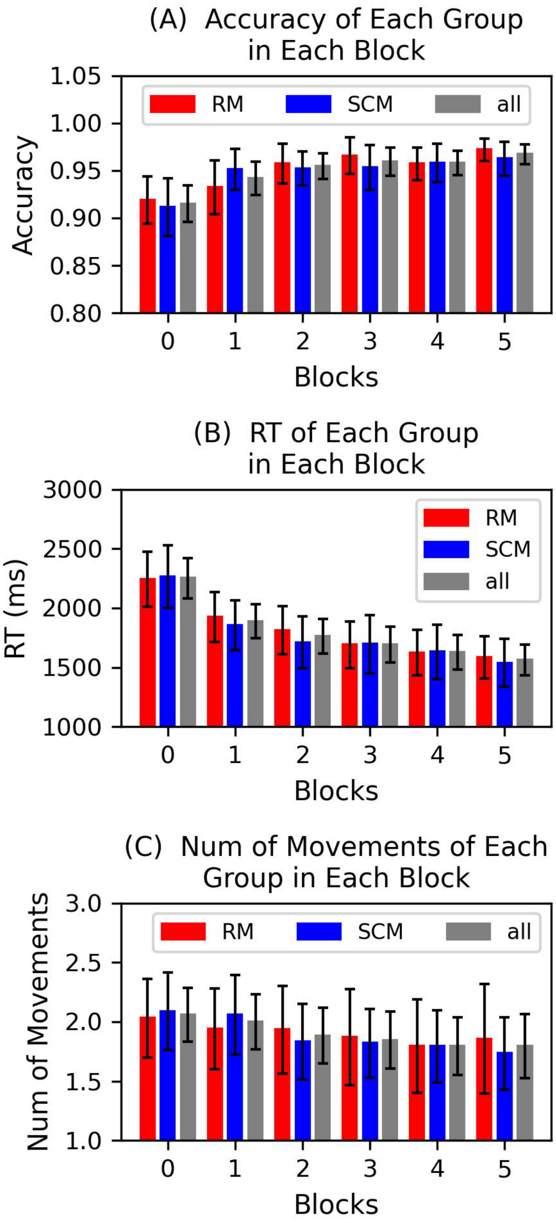


Figure 6.13. The action recognition accuracies (A), the action recognition RTs (B) and the numbers of within-trial view movements (C) in the 6 blocks of 32 trials of experiment 2.

There was a significant main effect of the block on the action classification accuracy, $F(3.992, 231.534) = 11.108$, $p < .001$, on the action classification RT, $F(2.506, 145.323) = 62.462$, $p < .001$, and on the amount of view movements, $F(2.487, 144.267) = 4.271$, $p = .010$. Figure 6.13(A) reveals that the accuracy in both groups improved almost at each block of trials, whereas Figure 6.13(B) shows the RT in both groups was shorter and shorter at each block. These results show a considerable learning effect since the participants in both groups processed more accurate and faster action recognitions over the whole test as they get more experience. Figure 6.13(C) shows that the number of viewpoint movements of all participants in both groups steadily reduced with more completed blocks.

Finally, none of the interaction effects of the block and the movement type on the action recognition accuracy, *F*(5, 300) = 1.016, *p* = .408, on the action recognition RT, *F*(5, 300) = .625, *p* = .681, and on the number of movements, *F*(5, 300) = .754, *p* = .584, were significant. These results suggest that the changes in accuracy, RT and number of movements across the blocks of trials were similar in both the RM and SCM groups. Briefly, the participants in both RM and SCM conditions showed the same learning effect.

6.3.2.3 Action Recognition Accuracies, Action Recognition RTs and the Numbers of View Movements in Each Action Class

The recognition of some action classes like dancing, sitting down, standing up and falling down from any viewpoint should be easier than discussing, pointing and waving which can only be spotted by a smaller set of viewpoints. Therefore, participants of all RM and SCM groups should have been able to recognise dancing, sitting down, standing up and falling down more accurately and more rapidly than discussing, pointing and waving. Furthermore, they should have done fewer view moves to recognise dancing, sitting down, standing up and falling down than discussing, pointing and waving.

I only calculated some descriptive statistics on the action recognition accuracies, the action recognition RTs and the number of view moves of the participants in each of the 7 action classes. No inferential statistics were made because that was not the purpose of the study and there was even the issue each participant was only tested on 4 random action classes from a pool of 7 action classes.

In this experiment with almost clear video images, the action classification accuracy overall was very high for all action classes. Anyway, the recognitions of the dancing (*M* = .977, *SD* = .034), sitting down (*M* = .959, *SD* = .054), standing up (*M* = .959, *SD* = .039) and falling down (*M* = .964, *SD* = .070) were slightly more accurate than discussing (*M* = .931, *SD* = .111), pointing (*M* = .900, *SD* = .115) and waving (*M* = .947, *SD* = .061). The action classification accuracies of all action classes are illustrated in Figure 6.14(A).

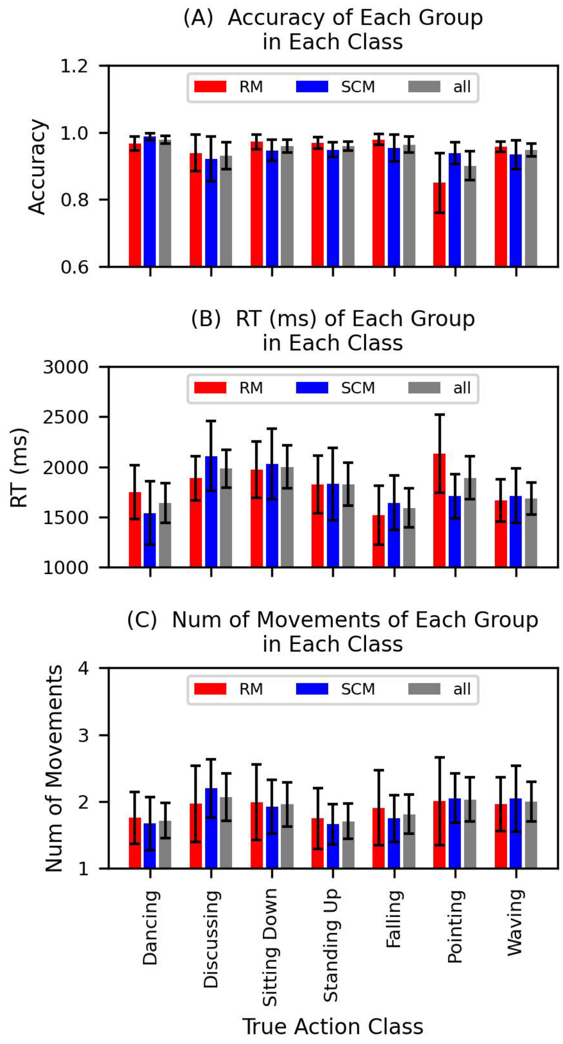


Figure 6.14. The action recognition accuracies (A), the action recognition RTs (B) and the numbers of within-trial view movements (C) in each of the 7 different action classes of the videos. These were the scores of experiment 2.

The classification of dancing (*M* = 1,642 ms, *SD* = 581 ms), falling down (*M* = 1,593 ms, *SD* = 559 ms) and waving (*M* = 1,687 ms, *SD* = 502 ms) were faster than discussing (*M* = 1,985 ms, *SD* = 522 ms), sitting down (*M* = 2,001 ms, *SD* = 609 ms), standing up (*M* = 1,828 ms, *SD* = 618 ms) and pointing (*M* = 1,893 ms, *SD* = 566 ms). In this experiment, the recognition of sitting down and standing up may have taken longer than the other classes because these actions actually appeared a little later in the videos than the other actions classes. The action classification RTs in all action classes are in Figure 6.14(B).

The participants classified dancing (*M* = 1.71, *SD* = .76), standing up (*M* = 1.71, *SD* = .76) and falling down (*M* = 1.81, *SD* = .84) with less viewpoint movements compared with discussing (*M* = 2.07, *SD* = .99), sitting down (*M* = 1.96, *SD* = .94), pointing (*M* = 2.03, *SD* = .88) and waving (*M* = 2.00, *SD* = .93). Figure 6.14(C) shows the numbers of view movements for all action classes.

The confusion matrix in Figure 6.15(C) shows that the participants confused pointing and waving with the largest confusion rate. The second and the third highest confusing rates were for discussing and pointing, and for discussing and waving, in turn.

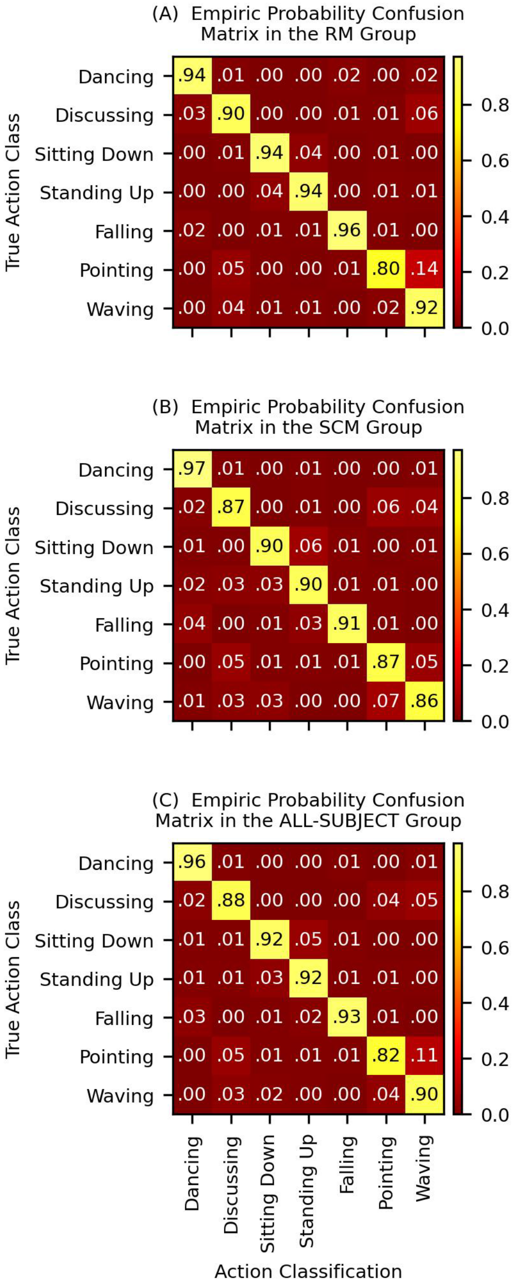


Figure 6.15. The confusion matrices of the RM (A), SCM (B) and all participants (C) in experiment 2.

6.3.2.4 Evidence of the Efficient and Inefficient Views from the Starting Views

6.3.2.4.1 Effect of the Starting View on the Accuracy, the RT and the Number of View Movements

I also compared the accuracies, the RTs and the quantity of view movements of all RM

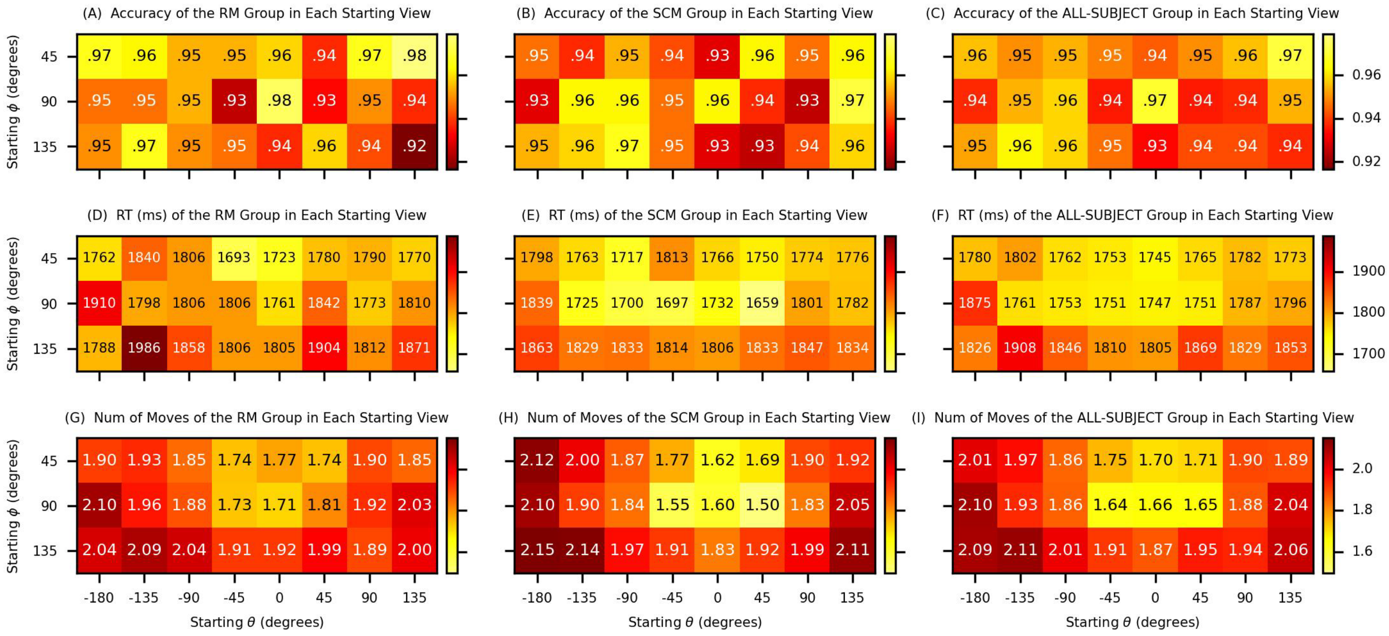


Figure 6.16. The accuracies of the RM (A), SCM (B) and all (C) participants in each starting view. The RTs of the RM (D), SCM (E) and all (F) participants in every starting view. The quantity of the view movements of the RM (D), SCM (E) and all (F) participants in every starting view. All these results come from experiment 2.

and SCM participants in all 24 starting view conditions by three one-way ANOVAs to explore the effect of the starting view on the action recognition accuracy, on the action recognition RT and on the number of view movements. Mauchly's tests showed that the assumption of sphericity had been violated for the accuracies, $\chi^2(275) = 361.623$, $p < .001$, the RTs, $\chi^2(275) = 392.742$, $p < .001$, and the view movements, $\chi^2(275) = 564.533$, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .652$ for the accuracy; $\varepsilon = .609$ for the RT; $\varepsilon = .498$ for the view movements). I did not follow up on any of the effect by t-tests because 276 pairs of 24 conditions with different starting views for each of the three dependent variables were too many.

No significant effect of the starting view on the accuracy was found, $F(14.993, 884.604) = 1.223$, $p = .248$. The ANOVA on the accuracies in the various starting views does not provide enough evidence of the difference between the efficient and inefficient views because the action recognition accuracy was not significantly impacted by the starting view. Figure 6.16(C) displays the accuracies of all participants of both groups given each starting view. However, there was a significant effect of the starting view on the RT, $F(14.006, 826.334) = 2.831$, $p < .001$. Since the action classification speed was affected by the starting view, we can reason the distinction between the efficient and inefficient views. Figure 6.16(F) illustrates the RTs of each starting view.

The starting view had a significant effect of on the numbers of within-trial view movements, $F(11.456, 675.925) = 10.056$, $p < .001$. The participants appeared to move their viewpoints away from the inefficient views towards the efficient ones because the view movements tended to be more numerous when their starting views were inefficient than when they were efficient. The heatmap in Figure 6.16(I) highlights the numbers of the within-trial view movements in each of the 24 starting views.

6.3.2.4.2 Effect of the Starting Angle ϕ on the Accuracy, the RT and Number of View Movements

Let’s now look at whether the starting angle ϕ effected the accuracy, the RT and the number of movements of both RM and SCM participants. Figure 6.17 shows the patterns of the accuracy, RT and the number of view moves of the RM, SCM and all participants as functions of the starting angle ϕ.

A one-way within-subject ANOVA was conducted to compare the effect of the starting angle ϕ on accuracy in top (ϕ_0=45°), middle (ϕ_1=90°) and bottom (ϕ_2=135°) starting views. There was not a significant effect of the starting angle ϕ on accuracy, $F(2, 118) = 1.136$, $p = .325$. This result suggests the starting angle ϕ did not have an effect on the accuracy of action recognition. Thus, it is not possible to identify the efficient and inefficient angles ϕ by just looking at the accuracy in all starting angles ϕ given that they were not significantly different.

A one-way within-subject ANOVA was conducted to compare the effect of the starting angle ϕ on RT in top (ϕ_0=45°), middle (ϕ_1=90°) and bottom (ϕ_2=135°) starting views. There was a significant effect of the angle ϕ on RT, $F(2, 118) = 14.373$, $p < .001$. Therefore, action recognition was faster in some starting angles ϕ than others. I followed up on this effect by running multiple paired t-tests with the Bonferroni correction to compare the RT in all 3 starting angles ϕ. The bottom views ϕ_2=135° (*M* = 1,843 ms, *SD* = 532 ms) were significantly slower than the top views ϕ_0=45° (*M* = 1,770 ms, *SD* = 537 ms), $t(59) = 4.98$, $p < .001$, and middle views ϕ_1=90° (*M* = 1,778 ms, *SD* = 532 ms), $t(59) = 4.18$, $p < .001$. The top views ϕ_0=45° were as fast as the middle views ϕ_1=90°, $t(59) = .515$, $p = 1.0$. These results suggest that the efficient angles ϕ were the top and the middle ones because participants were faster when they started the trials with either the top or middle views. On the other hand, the inefficient angle ϕ was the bottom one given that action recognition was slower when participants started watching from the bottom views.

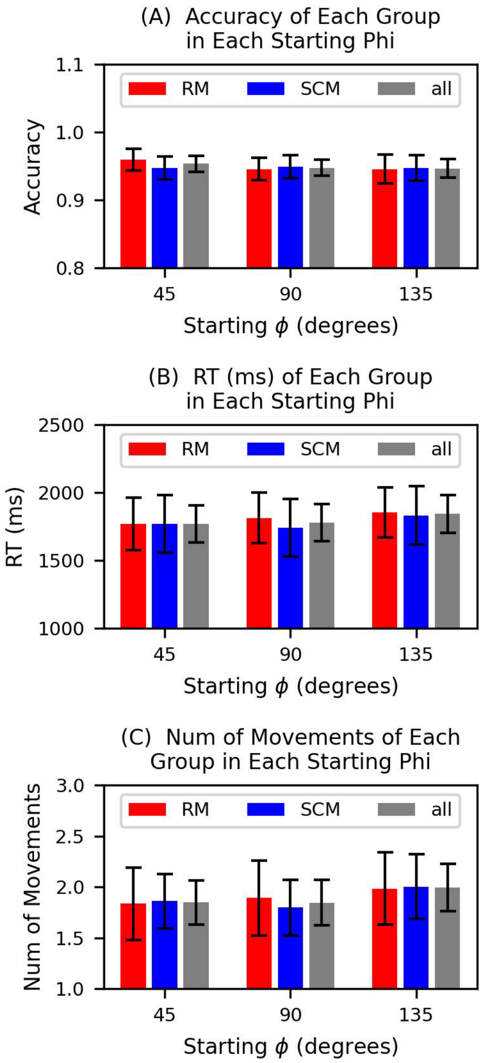


Figure 6.17. The accuracies (A), the RTs (B) and the quantity of the view movements (C) of the RM, SCM and all participants in each starting view angle ϕ. These results come from experiment 2.

Furthermore, a one-way within-subject ANOVA was conducted to compare the effect of the starting angle ϕ on the number of movements in top ($ϕ_0$=45°), middle ($ϕ_1$=90°) and bottom ($ϕ_2$=135°) starting views. There was a significant effect of the angle ϕ on the number of movements, *F*(2, 118) = 13.346, *p* < .001. I run multiple paired t-tests with Bonferroni correction to further investigate the latter effect on the number movements. Participants moved their view significantly more times if they started watching from the bottom views $ϕ_2$=135° (*M* = 1.994, *SD* = .896) than if they started watching from the top views $ϕ_0$=45° (*M* = 1.849, *SD* = .837), *t*(59) = 4.105, *p* < .001, and the middle views $ϕ_1$=90° (*M* = 1.845, *SD* = .866), *t*(59) = 4.301, *p* < .001. The participant with the top starting views approximately moved as many time as the middle starting view, *t*(59) = .139, *p* = 1.0. These results reveal a similar pattern that we saw on the RT in all starting angles ϕ. In fact, by looking at the number of movements, we can also tell that efficient views were the top and middles, while the inefficient views were the bottom views. That is because the participants moved less times when they started the trials from top and middle views, whereas they made more movement when they started the trial from the bottom views.

6.3.2.4.3 Effect of the Starting Angle θ on the Accuracy, the RT and the Number of View Movements

To highlight the efficient and inefficient angles θ, I also investigated the effect of the angle θ on the accuracy, RT and number of movements in all starting angles θ conditions. The overall patterns of the accuracy, RT and quantity of view movements as functions of the starting angle θ can be spotted in the charts of Figure 6.18 for each group of participants.

A one-way within-subject ANOVA was conducted to compare the effect of the starting angle θ on accuracy in all eight conditions. There was not a significant effect of the starting angle θ on accuracy, *F*(7, 413) = 1.026, *p* = .412. The accuracy did not vary, by varying the starting angle θ and did not highlight the efficient and inefficient angles θ.

A one-way within-subject ANOVA was conducted to compare the effect of the starting angle θ on RT in all eight conditions. There was a marginal effect of the starting angle θ on RT, *F*(7, 413) = 1.718, *p* = .103. I followed up on the marginal effect of the angle θ on RT with paired t-tests and Bonferroni correction. There was no significant difference between any pair of the eight conditions, mainly because there were 28 comparisons between eight conditions and the Bonferroni correction is extremely for such a huge number of comparisons. However, Figure 6.18(B) shows a slight and steady decrease of RT going from back to front starting views. Therefore, the efficient angles θ are the front ones because action recognition was slightly faster when the participants started watching from the front views. On the other hand, the inefficient angles θ are the back ones as the action recognition is slower if started from the back views.

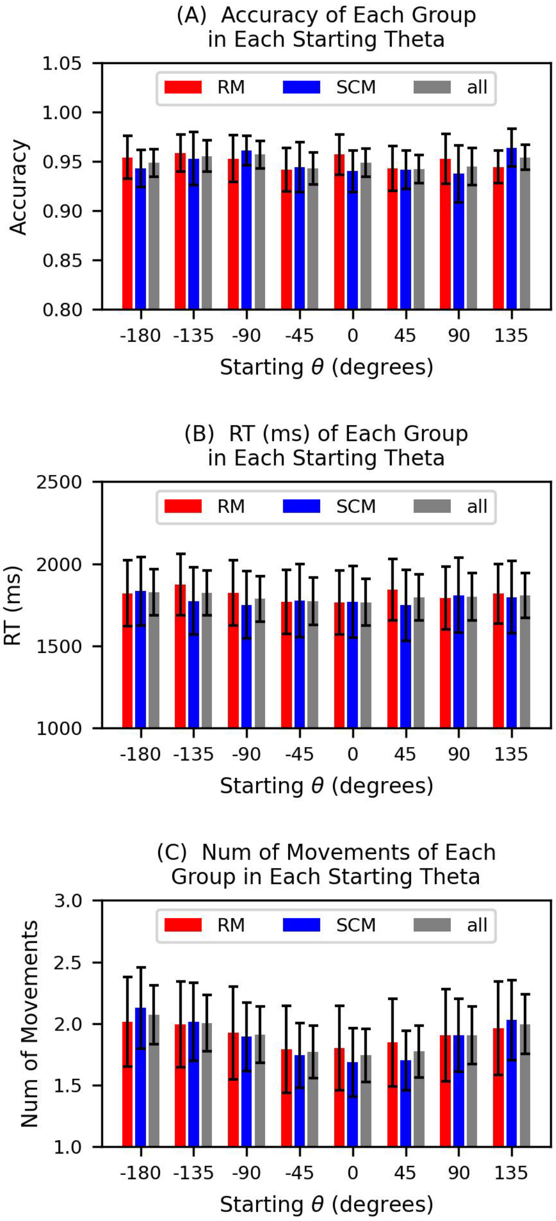


Figure 6.18. The accuracies (A), the RTs (B) and the quantity of the view movements (C) of the RM, SCM and all participants in each starting view angle θ. These were the results of experiment 2.

A one-way within-subject ANOVA was conducted to compare the effect of the starting angle θ on the number of movements in all eight conditions. Mauchly’s test indicated that the assumption of sphericity had been violated, $\chi^2(27) = 112.516$, $p < .001$. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .583). There was a significant main effect of the starting angle θ on the number of movements, $F(4.084, 240.983) = 20.525$, $p < .001$. I further investigated this effect with paired t-tests whose p-values were corrected by the Bonferroni method. Overall, these showed that participants significantly made more movements when they started watching from the back $\theta_0=-180°$ (M = 2.07, SD = .92), right-back $\theta_1=-135°$ (M = 2.004, SD = .88) and left-back $\theta_7=135°$ (M = 2.0, SD = .94) views than when they began watching from the right-front $\theta_3=-45°$ (M = 1.77, SD = .83), front $\theta_4=0°$ (M = 1.74, SD = .83), left-front $\theta_5=45°$ (M = 1.77, SD = .81) views. Again, the efficient angles θ seemed to be the front ones because participants moved less times if they started the trial from the front views. That is because they were already in a view with an efficient angle θ and did not need to move a lot. However, the inefficient angles θ were the back ones as they needed more view moves when they started watching from the back views. They might need more moves on their way from bad views to good views.

6.3.2.5 Evidence of Efficient View Selection

A goal of this study was to investigate whether the view selection of the SCM participants for action recognition is efficient. Hence, I looked at whether the human participants of the SCM condition tended to select the efficient views more frequently than the inefficient views.

6.3.2.5.1 View Selection

A one-way ANOVA was run on the percentage frequencies of all ending views of the SCM group to assess whether the SCM participants selected some views more than other views at the last timepoint of the video stimulation, just before the action recognition. Mauchly’s

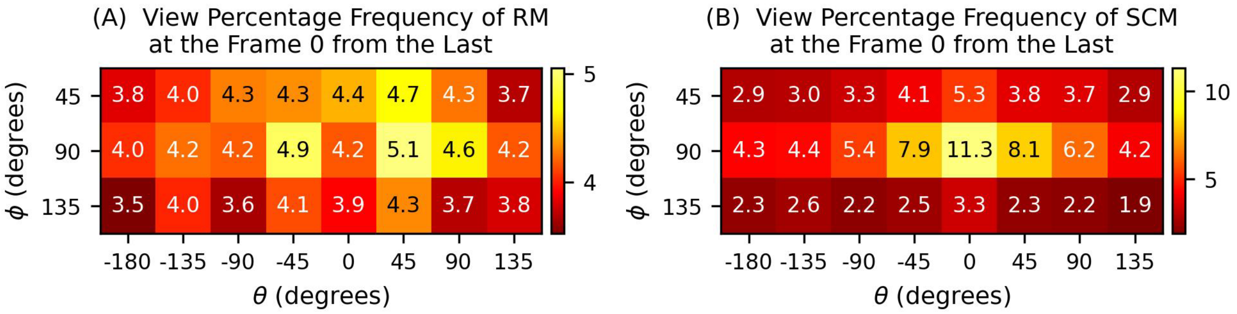


Figure 6.19. The percentage frequencies of the selected ending views in the RM and the SCM groups of experiment 2.

test indicated that the assumption of sphericity had been violated, $\chi^2(275) = 814.773$, $p <$ .001. Therefore, degrees of freedom were corrected by Greenhouse-Geisser estimates of sphericity ($\varepsilon = .145$). The percentage frequencies significantly varied across the different ending views, $F(3.328, 96.519) = 17.005$, $p < .001$. This result demonstrates that the SCM participants chose some views more often than others at the last timepoint of the video, just before their action classification. Figure 6.19(B) shows up the percentage frequencies of views at the last timepoint of all trials. Front views were chosen more frequently than the back and side views. Simultaneously, the top and bottom views were selected less often than the middle views.

6.3.2.5.2 Angle ϕ Selection

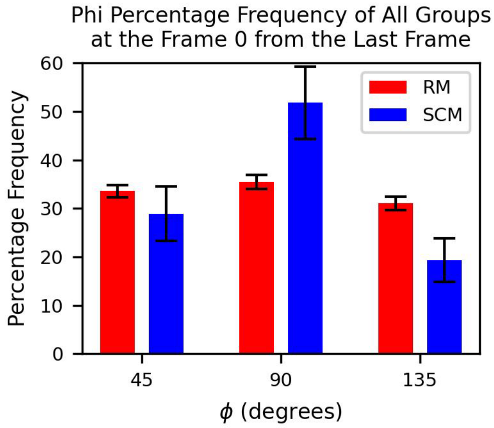


Figure 6.20. The percentage frequencies of the selected ending view angle ϕ in the RM and the SCM groups of experiment 2.

I ran an ANOVA to compare the percentage frequencies of ending angles ϕ in the SCM participants (Figure 6.20). The ending view angles ϕ could only be top, middle and bottom. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(2) = 12.064$, $p = .002$. Therefore, degrees of freedom were corrected using

Greenhouse-Geisser estimates of sphericity (ε = .741). The percentage frequencies of the ending angles ϕ were significantly different, *F*(1.481, 42.962) = 21.847, *p* < .001. I followed up on the effect by running multiple pairwise paired t-tests with the Bonferroni correction. The SCM participants significantly selected the middle angle ϕ (*M* = 51.782, *SD* = 19.940) more often than the top angle ϕ (*M* = 28.901, *SD* = 14.991), *t*(29) = 3.774, *p* = .002, and bottom angle ϕ (*M* = 19.317, *SD* = 11.908), *t*(29) = 6.084, *p* < .001. Likewise, the selection of the bottom angle ϕ was significantly less than the top angle ϕ, *t*(29) = −2.866, *p* = .023, and the middle angle ϕ, *t*(29) = −6.084, *p* < .001. In summary, the SCM participants selected more often the efficient angles ϕ which were the top and middle ones. They also selected less often the inefficient angle ϕ which was the bottom one.

6.3.2.5.3 Angle θ Selection

I then looked at whether the SCM participants selected more often the efficient angles θ than the inefficient angles θ of the views. Therefore, I conducted an ANOVA to compare the percentage frequencies of all eight ending angles θ in the SCM condition (Figure 6.21). Mauchly’s test indicated that the assumption of sphericity had been violated, $\chi^2(27)$ = 147.264, *p* < .001. Therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .311). The percentage frequencies of the ending angles θ were significantly different, *F*(2.175, 63.076) = 16.863, *p* < .001. I followed up on the effect with multiple pairwise paired t-tests between the percentage frequencies of the eight possible ending angles θ. The p-values were corrected by the Bonferroni correction. These

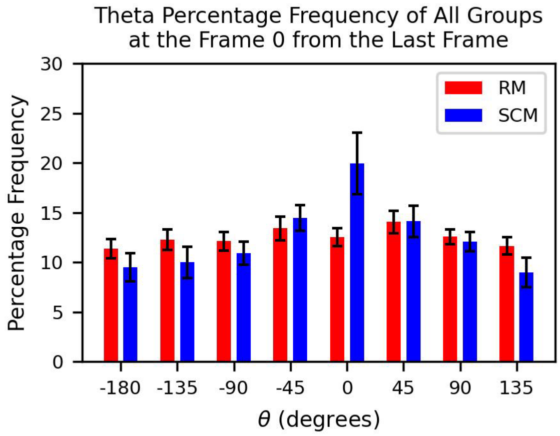


Figure 6.21. The percentage frequencies of the selected ending view angle θ in the RM and the SCM groups of experiment 2.

showed that participants significantly selected the right-front θ_3=−45° (*M* = 14.484, *SD* = 3.458), front θ_4=0° (*M* = 19.936, *SD* = 8.286), left-front θ_5=45° (*M* = 14.119, *SD* = 4.254) views more often than the back θ_0=−180° (*M* = 9.501, *SD* = 3.853), right-back θ_1=−135° (*M* = 9.988, *SD* = 4.177) and left-back θ_7=135° (*M* = 8.979, *SD* = 4.034) views. Thus, the participants in the SCM group selected the efficient angles θ more frequently than the inefficient angles θ to look at the actions.

6.3.2.5.4 Correlations Between the Selection of Views and the Action Recognition Performances of the same Views

A Pearson correlation coefficient was computed to assess the relationship between the percentage frequency of the ending views of the SCM group in the experiment 2 (Figure 6.19(B)) and the view accuracy of the NM group in the experiment of chapter 4 (Figure 4.2(A)). In the NM study, I excluded the left views and there were 15 views (5 angles θ x 3 angles ϕ), whereas, in this study, there were 24 (8 angles θ x 3 angles ϕ). By assuming that the right and left views were symmetric, I defined the accuracy in left-back views as the accuracy in the right-back views, the accuracy in the left views as the accuracy in the right views, and the accuracy in the left-front views as the accuracy in the right-front ones. In this way, I made 24 datapoints for the correlation: 24 view percentage frequencies and 24 view accuracies. The correlation matrix in Figure 6.22 indicate that there was a significant positive correlation between the two variables, *r*(22) = .6, *p* = .002. Overall, there was a strong, positive correlation between SCM view selection and the NM view accuracy. This result shows that SCM participants selected the views efficiently as they selected more often the efficient views with higher NM accuracy than the inefficient views with lower NM accuracy.

A Pearson correlation coefficient was computed to assess the relationship between the view percentage frequency in the SCM group in the Figure 6.19(B) and the NM view RT in the Figure 4.2(B) of chapter 4. Again, in the NM study, I excluded the left views and there were 15 views (5 angles θ x 3 angles ϕ), whereas, in this study, there were 24 (8 angles θ x 3 angles ϕ). By assuming that the right and left views were symmetric, I defined RTs in left-back, left, left-front views equal to the RTs in the right-back, right, right-front views, respectively. In this way, I made 24 datapoints for the correlation: 24 view percentage frequencies and 24 view RTs. As shown in Figure 6.22, there was a significant negative correlation between the two variables, $r(22) = -.62$, $p = .001$. This result suggests that SCM participants selected the views efficiently because they selected more often the efficient views with shorter NM RTs than the inefficient views with lower NM RT.

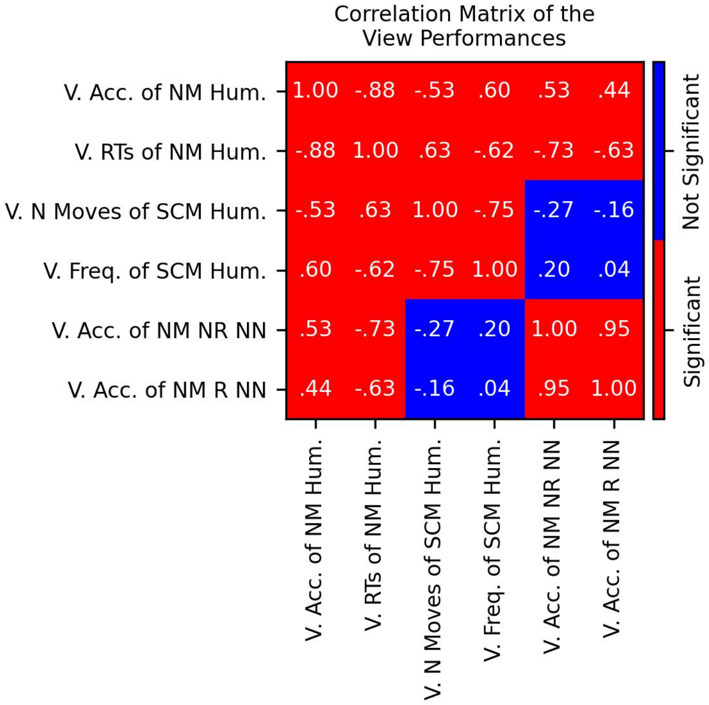


Figure 6.22. The correlation matrix of the following view performances: the view accuracies (V. Acc.) and the view RTs (V. RTs) of the human participants with no view motion (NM Hum.) from the study of the chapter 4; the numbers of view movements in each starting view (V. N Moves) and the percentage frequencies of the ending views (V. Freq.) in the SCM human participants (SCM Hum.) in experiment 2 of this chapter 6; the view accuracies (V. Acc.) of the non-recurrent (NR) and the recurrent (R) models (NN) of the study in chapter 5 with no viewpoint movements (NM).

6.4 Conclusions

The first objective of this chapter was to investigate whether the action recognition in the SCM condition was more accurate and faster than in the RM condition in the case of unclear images. Both online experiments of this chapter showed no significant differences in accuracy and RTs for the SCM and RM groups of human participants. I believe this was the case because both online studies had three methodological limits which may have added noise to both the accuracies and RTs of both the RM and SCM groups. The first limit was that the movement type was manipulated between subjects. Therefore, the differences between participants were not carefully controlled between the RM and SCM groups as much as they would have been if the movement type were manipulated within participants. The second limit was the fact that these were online studies. Hence, the participants were tested online far away from any experimenter. There would have been multiple advantages if they were tested in the lab with the presence of an experimenter: the participants would have been more motivated in doing the task well because they would have had the pressure that the experimenter would check whether their performance was good or bad; the experimenter would have checked whether the participants did the task properly and correct them; all participants would have used on the same PC. The third limit was the fact that four action classes were randomly chosen for each participant from a pool of seven action classes. This means that the selected action classes may have been unbalanced across the RM and SCM groups. In other words, the two groups may have not been tested on the same action classes. Therefore, given that some action classes can be recognised more accurately and more quickly than others and given that the two groups may have been tested on different action classes, then the results about the equality of the two groups in accuracy and RT may have been invalid.

Indeed, the action classification of SCM participants was 3.5% more accurate and 167 ms faster than one of the RM participants in the online experiment 1. Although these differences were not significant by the independent-sample t-tests, they are quite large. As far as I am concerned, they would have been significant by dependent-sample t-test if the view movement type, the independent variable, were manipulated within participants by a within-participant study design. As evidence, I also ran an independent-samples t-test and a paired-samples t-test on the RTs of the RM and SCM conditions of the pilot study, even if the view movement type was manipulated within participants. The difference in RT between the two conditions was a lot smaller (84 ms) in the pilot study than in the online experiment 1 of this chapter (167 ms). Yet, in the pilot study of chapter 3, the dependent t-test showed a significant difference in RT between the two conditions whereas the independent t-test showed no significant difference between them. Without going through some very mathematical explanations, the paired-samples t-test has more statical power than the unpaired-samples or independent-samples t-test.

The differences between the RM and SCM groups in accuracy and the RT were very insignificant in the online experiment 2. This may have been for the reasons above which are valid for both online experiments of this chapter. However, there is another reason that explains the insignificant differences in accuracy and RT between the two groups, specifically in the online experiment 2. This is the fact that the efficient view selection in the SCM condition for action classification may be less relevant if the images showing the actions are almost clear.

The second objective of the online studies of this chapter was to examine whether the SCM participants select the efficient views more frequently than the inefficient views even in the case of the more valid efficient and inefficient views of the chapter 4. Firstly, the frequencies of the selected ending views were significantly different in the SCM group of both online studies. Those results show that some views were selected more often than others. Critically, the frequencies of the selected ending views of the SCM group of both online studies of this chapter 6 were positively correlated to the accuracies of the locked views in the study of chapter 4 and negatively correlated to the RTs of the locked views in the same study of chapter 4. These other results highlighted that the SCM human participants selected the efficient (more accurate and faster) views more often than the inefficient (less accurate and slower) views, even in the case of the more valid efficient and inefficient views of the chapter 4.

Chapter 6 aimed to replicate the results of chapter 3 proving the efficient active action classification of the human observers in the case of different methods. There was a replication of the human observers selecting the efficient views more often than the inefficient views for active action classification. However, the studies of chapter 6 did not replicate the higher action recognition performances of the SCM participants over the RM participants. In my opinion, if I do the same studies in the lab, if I manipulated the view movements type (RM vs SCM) within participants and if I show the exact same actions in the RM and SCM conditions, then I would get the expected higher performance in the SCM condition with respect to the RM baseline condition.

7 Active Action Recognition of the Robotic Observers with Supervised Learning and Deep Q-Learning

7.1 Introduction

Active Robotic vision (or active computer vision) is a discipline of computer vision that designs and studies active vision models that analyse visual data such as images and videos to select their next best views and ultimately improve their performance in other visual tasks. Some instances of these other visual tasks are object classification, object tracking, object manipulations, body pose estimation, action classification, and agent interaction. Some good examples of studies about active computer vision are (Arzati & Arzanpour, 2021; Jayaraman & Grauman, 2018; Ramakrishnan & Grauman, 2018; Roost et al., 2020). Generally, these studies use deep Q-learning (DQN) to train some active NNs in selecting the next best views. The DQN is a reinforcement learning technique that applies q-learning to NNs. This chapter focuses on the active action classification of robotic observers which analyse the video images to select the next best views and increase their performance in classifying the actions of human actors in the observed images.

The study of Jayaraman and Grauman (2018) showed that an active vision NN intelligently selected the viewpoints to visually analyse some objects and improved their performance in the visual perspective taking task. The visual perspective taking task was to reconstruct (or predict) the observed images of some objects from a few viewpoints, as well as the unobserved images of the same objects from many other viewpoints. They rendered 84 images for every 3D object of the dataset ModelNet (Wu et al., 2015) from 84 different view positions (12 angles θ x 7 angles ϕ). Then, the observer's task was to predict all 84 images of an object using a limited number of input images, which were far fewer than the total number of views. To accurately predict the unseen views, the observer had to navigate around the object intelligently and select the most informative views. They trained the system to select the next-best view using reinforcement learning, where the reward was the model's own negative image prediction loss. The active model outperformed all baseline passive models, including the one-view model and the random view movement model, showing that it had learned to select the most informative views to construct 3D representations of objects. The one-view model predicted all 84 images of an object by observing that object from only one viewpoint. The random view movement model predicted the views by observing the objects from as many viewpoints as in the active model. However, the next view was selected randomly for the random model and not chosen by the model itself. Additionally, by framing the task as visual perspective taking rather than object recognition, the authors managed to train the models to construct 3D representations of objects at a low cost, as the data was label-free.

This chapter aimed to investigate whether robotic observers can learn by DQN to do efficient active action classifications. Therefore, I ran the active computer vision experiment in this chapter. All models in the experiment were RCNNs and their architectures were exactly the same. Their architecture included some convolutional layers, an LSTM (Hochreiter & Schmidhuber, 1997) layer, and some fully-connected layers. At each timepoint of a trial, these models classified the actions of the actors in the images and selected the view for the next timepoint. I discussed in more detail in chapters

1 and 4 the reasons why RCNNs are the most appropriate models for active action recognition.

I trained all models with active or passive view movements to select the next best views by the DQN algorithm. The rewards for the models at any timepoints were determined by the negative values of the action classification losses at the next timepoints. The models were trained to minimise the reward prediction losses for the selected view movements, regardless of whether the movements were chosen actively or passively. The reward prediction losses measured the discrepancy between the model's predicted rewards for the selected view movements and the actual rewards that resulted from those movements. The reward prediction losses were used as an evaluation metric to improve the efficient view selections of all models during their training with active or passive view moves. Once these efficient view selections were trained, these were expected to ultimately lead to higher rewards and more accurate action classifications if the models were then tested with active view movements than with passive view movements.

There were active and passive view movements in this experiment. On one hand, the active view movements were selected by the models themself. I will refer to these active movements also as SCMs. On the other hand, passive view movements were not chosen by the models themself. This study used two types of passive view movements: the RMs and the NMs. The models with NMs classified the actions from some predefined viewpoints whose positions were locked and did not change over time. The models with RMs classified the actions from viewpoints that changed randomly. The RMs were passive because they were not selected by the models. The models with NMs and RMs also classified the actions and selected the next viewpoint movement at each timepoint. However, the selected view movements of the models with NMs and RMs at each timepoint were ignored and replaced with either some NMs or some RMs, respectively.

There were four groups of models in this study. The models of different groups were trained, validated and tested with different types of view movements. Two groups were passive and the other two groups were active. The two passive groups were the NM and RM groups. The NM models were trained validated and tested with only NMs, while the RM models were trained, validated and tested with only RMs. On the other hand, the other two active groups were the SCM and the RSCM groups. The SCM models were trained with both RMs and SCMs. However, these were validated and tested with only SCMs. Lastly, the random and self-controlled view movement (RSCM) models were trained and validated with only RMs and tested with only SCMs.

This study had two specific objectives. The first one was to examine whether the action recognition of any active group was more accurate than all passive groups, once they had been trained to select the next best views by DQN and to classify the actions in the dynamic videos. Therefore, either the SCM models or RSCM models were expected to classify the actions in the dynamic videos more accurately than both the NM models and the RM models. If the action recognition of any group of the active models were indeed more accurate, then this result would be evidence that the robotic observers can learn to process efficient active action classifications by DQN.

The second objective was to inspect whether the active action classifiers selected their efficient views for their action recognition more frequently than their inefficient views when they had been trained by DQN. I exposed the efficient and inefficient views for the action recognition of the robotic observers with locked viewpoints in chapter 5. In the study of this chapter, the efficient and inefficient views for the robotic observers could also be highlighted by identifying the views from where the passive NM models were more accurate and the views from where they were less accurate. If my active models selected the efficient view more often than the inefficient ones, then the percentage frequency of the selected views of the active models at the last timepoints should have been positively correlated to the action recognition accuracies of all views which were scored by the passive models of chapter 5 and by the NM models of this chapter 7. This would be another proof that robotic observers can compute efficient active action classifications, by using DQN.

7.2 The Architecture of the Active Action Classifiers

The architecture was the same for all active action classifiers. Each model consisted of three parallel sub-models which were an action classifier and two view movement selectors. The task of the action classifier was to classify the actions in the videos. The two view movement selectors were the angle θ movement selector and the angle ϕ movement selector whose tasks were to select the changes of the current view angles θ and the changes of the current view angles ϕ, respectively, that would ultimately lead to more accurate action classifications.

The two view movement selectors were DQN policies whose tasks were to predict the rewards at the future timepoints for all allowed view movements given the environment states at the current and previous timepoints. Since the two view movement selectors contained a recurrent layer, an LSTM, that integrated the observations across time, the environmental states of the two policies at the timepoint t were the observed video images at the timepoints from 0 to t. The actual reward at the timepoint t was the negative action classification loss at the next timepoint t+1. Once the view movement selectors had been trained to predict the future rewards for each view movement given the environmental states, they could select the view movements with the largest negatives of predicted action classification losses or the smallest positives of predicated action classification losses. In other words, by choosing the view movements at the timepoint t with the largest predicted rewards, the view movement selectors selected the view movements with the highest negatives of predicted action classification losses or the smallest positives of action classification losses.

The architectures of the three sub-models were identical, except for the last layer whose size was 3 for each of the two view movement selectors and 7 for the action classifiers. Each sub-model of active action classifiers had a total of 20 layers. The first 18 layers were shared by the three sub-models, whereas the last two layers were segregated. The first 17 shared layers were a 2d CNN feature extractor which was a finetuned 18-layer ResNet without the last classifying layer. The input of the feature extractor was a batch of B x T raw coloured images with the size of 3 x 224 x 224, where B and T could be any positive integers and were the number of videos and the number of images per video, in turn. The output of the extractor was an array of B x T x 512 image features, where 512 was the number of features per image. The remaining shared layer was an LSTM with both weights and biases. The LSTM’s input was the output of the feature extractor and there were B x T x 1,000 image features in the output of the LSTM.

The last stack of two layers of each sub-model was independent of (not shared by; parallel to) the other sub-models. Both layers of each sub-model were fully-connected with both weights and biases. The first independent layer of each sub-model fed on the LSTM’s output and returned B x T x 200 features. These were then fed to the second layer of the same sub-model, whose output was either an array of B x T x 3 reward predictions in the case of each view movement selector or B x T x 7 action class predictions in the case of the action classifier.

7.3 Experiment

7.3.1 Dataset

The goal was to create the dynamic environments of the models by the MVVs of the robotic version r.3.5 of the MVVHA. Each dynamic environment or dynamic video was the observed images which were chosen from the one MVV according to some view movement rules. However, I did not directly use this dataset version with raw images. I only used the S=10 feature datasets which were produced in the experiment of chapter 5. Each of these was extracted by one of the S non-recurrent feature extractors of chapter 5 from the raw images in the robotic version r.3.5 of the MVVHA. These extractors were finetuned 18-layer ResNets without the last classifying layer. They were only finetuned to classify the actions in raw images and not to select the next best views. The difference between the S feature extractors was only their finetuning experience since each of them was finetuned with different training actors.

7.3.2 Environment

The environments of the models were the dynamic videos whose image features were sampled from the MVVs of the feature datasets. The models observed the 2-second dynamic videos at 3Hz since they only observed T=6 frames out of 60 per dynamic video of all feature datasets. These videos were dynamic because their viewpoints could move at any timepoint according to some movement rules. The view of the first timepoint was passively sampled from the pool of 40 views in each feature dataset. However, the view at any timepoint between the second and the last ones was defined by the view and the view movement of the previous timepoint. One of the nine possible view movements was passively or actively chosen at each timepoint to define the view of the next timepoint. The selected movement at any timepoint could move the observer's view at the same timepoint to one of its nine neighbouring views. The combination of three possible θ movements and three possible ϕ movements defined these nine possible view changes.

At each timepoint, one of the three possible θ movements and one of three possible ϕ movements were passively or actively chosen to define the view of the next timepoint. The view angle θ_{t+1} for the next timepoint t+1 was a function of the view angle θ_t at the tth timepoint and the selected movement $\Delta_{0,t}$ of the angle θ_t at the tth timepoint, as in equation 7.1.

$$\theta_{t+1} = \theta_t + \Delta_{0,t} \qquad 7.1$$

The selected movement $\Delta_{0,t}$ was sampled from a set of three possible θ movements. The first possible movement $\Delta_{0,t} = -45°$ was to the left neighbouring view, the second $\Delta_{0,t} = 0°$ was a no movement, and the third $\Delta_{0,t} = +45°$ was to the right neighbouring view.

Similarly, the view angle ϕ_{t+1} for the timepoint t+1 was defined by the view angle ϕ_t at the tth timepoint and the selected movement $\Delta_{1,t}$ of the angle ϕ_t at the tth timepoint, as equation 7.2 shows.

$$\phi_{t+1} = \phi_t + \Delta_{1,t} \qquad 7.2$$

The selected movement $\Delta_{1,t}$ was sampled from a pool of three possible ϕ movements. The first possible movement $\Delta_{1,t} = -45°$ was to the upper neighbouring view, the second $\Delta_{1,t} = 0°$ was a no movement, and the third $\Delta_{1,t} = +45°$ was to the lower neighbouring view.

At each timepoint, the task of the action classifier of the models was to classify the action in the video, while the tasks of the θ movement selector and the angle ϕ movement selector were to select one of the three possible θ movements and one of the three possible ϕ movements, respectively. The two movement selectors selected the θ and ϕ movements they predicted to obtain the highest rewards given the current observation.

The selected movements at any timepoint could be either SCMs, NMs or RMs. The SCMs were active because they were the selected movements of the two movement selectors of the models. However, the NMs and the RMs were passive because the selected movements of the two movement selectors were ignored and replaced with no view movements and some random view movements, respectively.

7.3.3 Four Groups of Models

There were four groups of S=10 models in the experiment of this chapter. Two groups out of four were active whereas the other two ones were passive. There was one group of ten SCM models, one of ten NM models, one of ten RM models, and another one of ten RSCM models. The SCM and RSCM groups were active whilst the NM and RM groups were passive. The SCM group was active because the models of this group were trained with passive RMs and active SCMs and then tested with only active SCMs. The probability ε_e of RMs was defined by the equation 7.3, while the probability of SCMs was $1-\varepsilon_e$. In the first training epoch of the SCM models, the probability ε_0 of RMs was 1 and the probability $1-\varepsilon_0$ of SCMs was 0. From the second epoch on, the probability ε_e of RMs at the eth epoch was 60% of the previous ε_{e-1} at the epoch e-1. If ε_e was less than 0.05, then ε_e was 0.05.

$$\varepsilon_e = \begin{cases} 1 & if\ e = 0 \\ 0.60 \times \varepsilon_{e-1} & if\ \varepsilon_e \geq 0.05 \\ 0.05 & otherwise \end{cases} \qquad 7.3$$

The NM models were passive since they were trained and tested with only passive NMs. Similarly, the RM models were also passive, given that they were trained and tested with only passive RMs. The RSCM models were also active because they were trained with only passive RMs and then tested with only active SCMs. Indeed, the RSCM models were the trained RM models which were tested with SCMs. The RSCM models could still learn to predict the rewards and select the next best views even if they were trained with the RMs by utilising the reward prediction losses of the randomly selected view movements. Since the analyses were only run on the models' performances of the test rather than their performances of the training and given that the RSCM models were tested with only active SCMs, then the RSCM models were considered to be another active group.

Every model of this study did not actually include a feature extractor, because each one was neither trained, validated or tested on raw images but on the image features of the feature datasets. Because the feature datasets were already the output of feature extractors, the actual models of this study did not include the first 17 layers which were defined by the finetuned 18-layer ResNets without the last layer and they only included the last three layers from the 18th layer (the LSTM) to the last one.

7.3.4 Multi-Task Loss

The multi-task loss l for any batch of B dynamic videos with lengths of T frames was defined by equation 7.4. The terms l_{slc_0}, l_{slc_1} and l_{cls_0} are the single-task losses of the angle θ movement selector, the angle ϕ movement selector and the action classifier, respectively. The terms w_{slc_0}, w_{slc_1} and w_{cls_0} are the weights of these single-task losses, in turn.

$$l = w_{slc_0} l_{slc_0} + w_{slc_1} l_{slc_1} + w_{cls_0} l_{cls_0} \qquad 7.4$$

The multi-task loss *l* was more generally defined by the equation 7.5 for any F movement selectors and for any C classifiers, where F and C are positive integers. In this equation, l_{slc_f} and w_{slc_f} are the single-task loss and loss weight of the fth movement selector, while

l_{cls_c} and w_{cls_c} are the single-task loss and loss weight of the cth classifier, respectively.

$$l = \sum_{f=0}^{F-1} w_{slc_f} l_{slc_f} + \sum_{c=0}^{C-1} w_{cls_c} l_{cls_c} \qquad 7.5$$

The sum of the loss weights was 1, as in equation 7.6.

$$\sum_{f=0}^{F-1} w_{slc_f} + \sum_{c=0}^{C-1} w_{cls_c} = 1 \qquad 7.6$$

The single-task loss l_{cls_c} was the average of the classification losses L_c of the cth classifier given a batch of B dynamic videos with length T. The equation 7.7 calculated the single-task loss l_{cls_c}.

$$l_{cls_c} = \frac{1}{B \times T} \sum_{b=0}^{B-1} \sum_{t=0}^{T-1} L_{c,b,t} \qquad 7.7$$

$L_{c,b,t}$ was the classification loss of the cth classifier for the observation at the tth timepoint in the bth dynamic video in the batch. Since the cth classifier was recurrent, its observation at the tth timepoint in any dynamic video comprised the first t+1 images of that video at timepoints from 0 to t. The classification loss $L_{c,b,t}$ was determined by the cross entropy loss (Mao et al., 2023; Zhang & Sabuncu, 2018), as it is shown in equation 7.8, where $z_{c,b,t}$ is the vector of K_c class predictions (before softmax) of the cth classifier given the observation at the tth timepoint in the bth dynamic video, K_c is the number of all possible classes for the cth classifier, and $y_{c,b,t}$ is any integer in the range $0 \leq y_{c,b,t} < K_c$ and represents the true class that the cth classifier attempted to predict given the observation at the tth timepoint in the bth dynamic video.

$$L_{c,b,t} = -\log\left(softmax(z_{c,b,t})\right)_{y_{c,b,t}} \qquad 7.8$$

The equation 7.9 defines the softmax function.

$$softmax(z_{c,b,t,k}) = \frac{\exp(z_{c,b,t,k})}{\sum_{i=0}^{K_c-1} exp(z_{c,b,t,i})} \tag{7.9}$$

The single-task loss l_{slc_f} was defined in equation 7.10 as the average of the reward prediction losses L_f of the fth movement selector given a batch of B dynamic videos with length T. The reward prediction loss $L_{f,b,T-1}$ at the last timepoint of the bth video was excluded from the average simply because it was not available. The reason is that the reward prediction loss $L_{f,b,t}$ was a function of the reward $r_{b,t}$ and the next expected return $V_{slc_f}(x_{b,t+1})$, which could not be defined for the last observation as well, because they were functions of the next observations and there were no next observations for the last observation T-1.

$$l_{slc_f} = \frac{1}{B \times (T-1)} \sum_{b=0}^{B-1} \sum_{t=0}^{T-2} L_{f,b,t} \tag{7.10}$$

The reward prediction loss $L_{f,b,t}$ was calculated by the smooth L1 loss of the temporal difference error $\delta_{f,b,t}$, as in equation 7.11.

$$L_{f,b,t} = smooth_{L_1}(\delta_{f,b,t}) \tag{7.11}$$

The smooth L1 loss (Girshick, 2015) is defined in equation 7.12.

$$smooth_{L_1}(\delta_{f,b,t}) = \begin{cases} 0.5 \times \delta_{f,b,t}{}^2 & if|\delta_{f,b,t}| < 1 \\ |\delta_{f,b,t}| - 0.5 & otherwise \end{cases} \tag{7.12}$$

However, the temporal difference error $\delta_{f,b,t}$, was computed by the equation 7.13.

$$\delta_{f,b,t} = Q_{slc_f}(x_{b,t}, a_{f,b,t}) - \left(r_{b,t} + \gamma \times V_{slc_f}(x_{b,t+1})\right) \tag{7.13}$$

The term $Q_{slc_f}(x_{b,t}, a_{f,b,t})$ was the action-value (reward prediction) for the actively or passively selected action $a_{f,b,t}$ given the environment state (observation) $x_{b,t}$ at the tth timepoint of the b^th^ dynamic video according to the f^th^ policy (movement selector) slc_f. The term $r_{b,t}$ was the actual reward for the f^th^ policy slc_f that resulted from the selected action $a_{f,b,t}$ at the state $x_{b,t}$. The discount factor γ of the future rewards was set to .99 for all F movement selectors of all models in the whole experiment. Instead, $V_{slc_f}(x_{b,t+1})$ was the expected return (reward prediction) given the next state $x_{b,t+1}$ at the next timepoint t+1 of the same b^th^ video that resulted from the selected action $a_{f,b,t}$ at the state $x_{b,t}$, if all future actions from timepoint t+1 on were actively selected by the same f^th^ policy slc_f.

The expected return $V_{slc_f}(x_{b,t+1})$ was the maximum of the action-values for all A_f possible actions of the f^th^ movement selector slc_f, as in the equation 7.14.

$$V_{slc_f}(x_{b,t+1}) = \max_{a_f \in A_f}\left(Q_{slc_f}(x_{b,t+1}, a_{f,})\right) \qquad 7.14$$

Finally, the equation 7.15 shows that the reward $r_{b,t}$ was defined by the negative weighted average of the single-task classification losses of all C classifiers given the next observation at timepoint (t+1) of the same video. Additionally, the reward $r_{b,t}$ was biased by a constant $bias_r$. This constant was not a trainable parameter of the models. In fact, it was set to 1 and remained the same in the whole experiment.

$$r_{b,t} = -\frac{\sum_{c=0}^{C-1}\left(w_{cls_c} L_{c,b,t+1}\right)}{\sum_{c=0}^{C-1} w_{cls_c}} + bias_r \qquad 7.15$$

This multi-task loss is similar to the ones used in object detection (Girshick, 2015; Redmon et al., 2016) where models learned to estimate the classes of the objects in images and the parameters of bounding boxes containing these objects. The parameters of the bounding boxes were the x and y coordinates of the object centres, the width and the height of the objects. Predicting the parameters of the bounding boxes was a regression task while classifying the objects was a classification task. Since regression and classification require different type of single-task loss functions, they used a multi-task loss function to train the models for object detection. The multi-task loss was the sum of the scaled single-task loss functions of the model's prediction of the bounding box parameters and the model's classification of the objects.

7.3.5 Multiple dataset splits

Similarly to the study in chapter 5, there were S=10 models per group in this study instead of only one model per condition to more statistical power of the results. The only difference between the models of the same groups was their training experience given that they were trained with different videos showing different actors, like in the experiment of chapter 5. I even used the same S splits of 5 training actors, 10 validation actors, and 50 test actors as in chapter 5.

The S feature datasets and the S dataset splits were the same between the four groups of S models. However, each model was trained, validated and tested with a different feature dataset and a different dataset split from the other S-1 models of the same group. Formally, the s^{th} model of the g^{th} group was assigned to the s^{th} feature dataset with the s^{th} dataset split. The s^{th} split of the s^{th} model was the same as the s^{th} split of the s^{th} feature extractor by which the s^{th} feature dataset was extracted from the raw image dataset. This was done to keep continuity between the two parts of the architecture, even if they were trained at different moments.

7.3.6 Two Experiment Phases

There were two experiment phases. In each phase, all models were first trained and then tested. There were only two differences between the two phases: The first one concerned the model parameters (weights and biases). Completely untrained models were initiated at the beginning of the training in the phase, while the semi-trained models, which had already completed the training of the first phase, started the training of the second phase. Therefore, the model parameters were completely untrained at the start of the first phase, while they were semi-trained at the beginning of the second phase.

The second difference between the two experiment phases was in the single-task loss weights w_{slc_0}, w_{slc_1} and w_{cls_0} in the equations 7.4, 7.5 and 7.6. The weights w_{slc_0}, w_{slc_1} and w_{cls_0} were set to 0.25, 0.25 and 0.50 in the first phase, respectively. Nevertheless, the weights w_{slc_0}, w_{slc_1} and w_{cls_0} were equal to .40, 0.40 and 0.20 in the second phase, in that order. Thus, each weight of the movement selectors was half as much as the weight of the action classifier in the first phase, whereas each weight of the movement selectors was double the weight of the action classifier in the second phase.

Everything else was the same in the two phases. For instance, while both the feature datasets and the dataset splits were different between the models of the same group, both were the same between the two experiment phases of the same model.

7.3.7 Code and Hardware

The whole experiment, including the training and the testing of the active and passive action classifiers, was scripted with python code and some python libraries like PyTorch (Paszke et al., 2019), NumPy (Harris et al., 2020) and CalaPy (pypi.org/project/calapy). CalaPy is my python library that I fully coded.

I ran the training and the testing of all models in the two phases in parallel on the 24 high-performance NVidia GPU cards of the Ceres cluster. The model of 16 GPU cards was GTX1080Ti, and the model of the other 8 GPU cards was RTX2080. The Ceres cluster is a computational cluster comprising 2192 processors (including multi-threading), 43.5Tb total RAM, 24 GPUs, and 660Tb of dedicated storage.

7.4 Results

7.4.1 Effect of Phase and Time

The overall action recognition accuracy of all groups of models was higher in the second experimental phase (phase 1) than in the first one (phase 0). This pattern is highlighted in Figure 7.1. Additionally, the action recognition accuracy of all groups of models peaked at the fifth timepoint (timepoint 4). Figure 7.2 shows the accuracy of each model group at each timepoint. Therefore, I am only going to show in this chapter the models' results at timepoint 4 in phase 1.

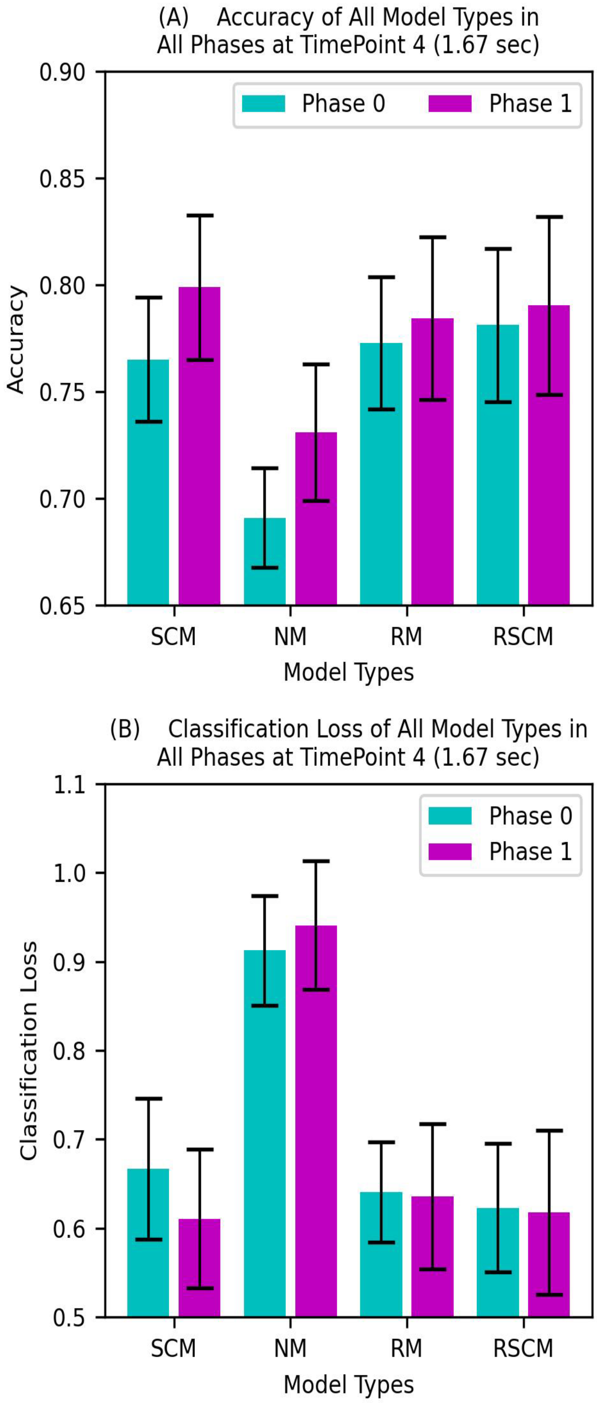


Figure 7.1. The action classification accuracy (A) and the action classification loss (B) of each group of models in each experiment phase.

7.4.2 The Active Models were More Accurate than the Passive Models

Figure 7.1 displays the action recognition accuracies of the four groups of models. The NM models were the fourth most accurate (*M* = 0.731, *SD* = 0.045), the RM models were the third (*M* = 0.784, *SD* =0.053), the RSCM models were the second (*M* = 0.790, *SD* =0.058) and the SCM models were the first (*M* = 0.799, *SD* = 0.047). Thus, the RM models were 0.053 (5.3%) more accurate than the NM models, showing that the RMs improve the action recognition of the models even if these movements are passive. In addition, both the active SCM and RSCM models were more accurate than the passive RM models by 0.015 (1.5%) and 0.006 (0.6%), in that order. This proves that our active models efficiently selected the next best view to improve the accuracy of their action classification.

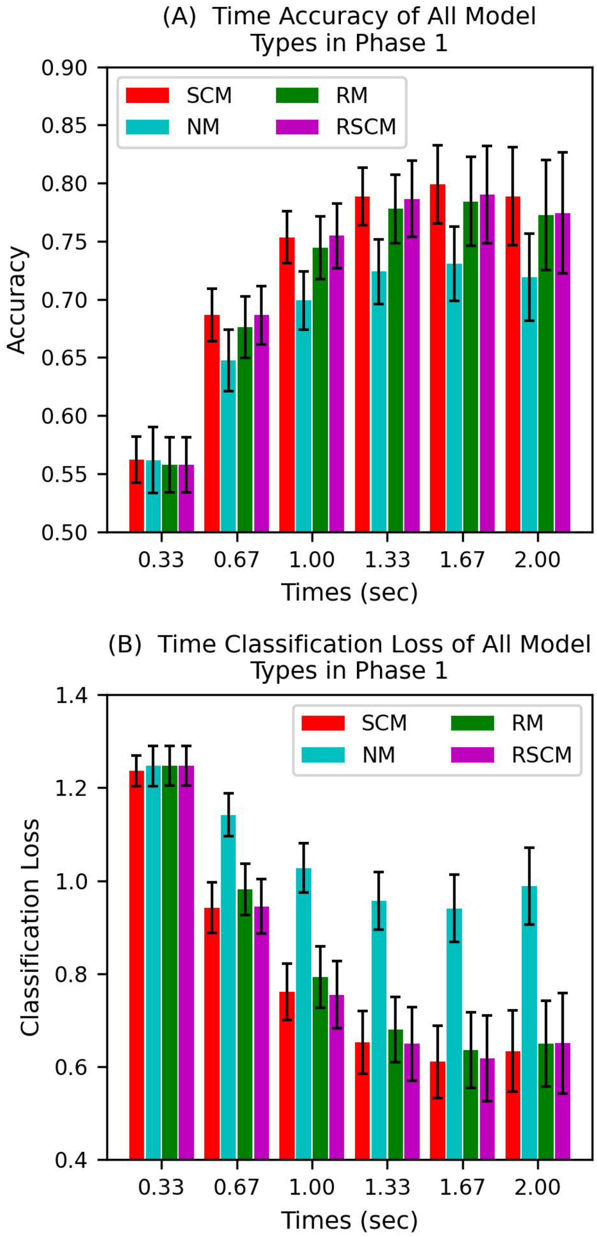


Figure 7.2. The action classification accuracy (A) and the action classification loss (B) of each group of models at each timepoint of the second experiment phase.

7.4.3 Replication of the Efficient and Inefficient Views for the Action Recognition of the Robotic Observers

The efficient and inefficient views for the passive NM models of this study (Figure 7.3) were identical to the ones for the passive recurrent models of the study in chapter 5 (Figure 5.7). Figure 7.3 highlighted three main results for the NM models which were also valid for the recurrent models of chapter 5. First, the accuracies of the NM models from the left views (views on the left side of the actors) with θ=−45°, θ=−90° and θ=−135° were symmetric with the accuracies of the same models from the right views (views on the right

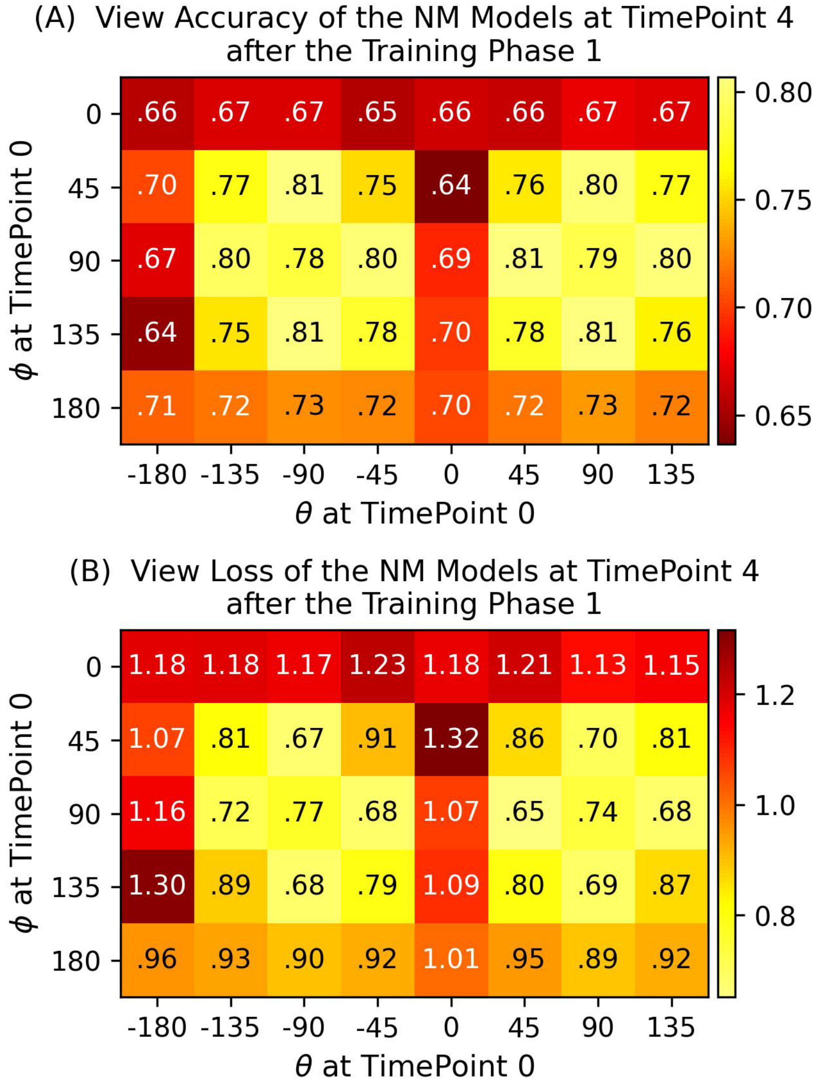


Figure 7.3. The action classification accuracy (A) and the action classification loss (B) of the NM group of models from each viewpoint at the fifth timepoint of the second experiment phase.

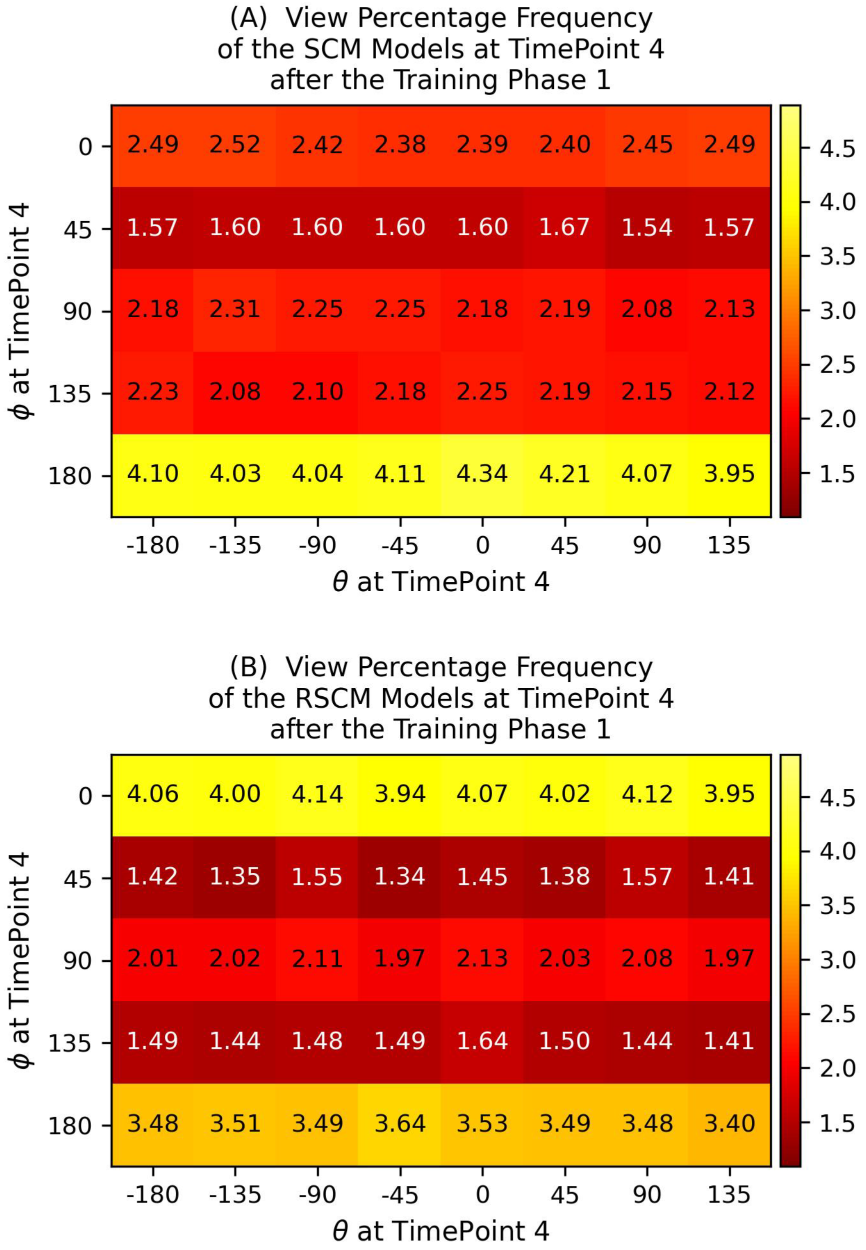
side of the actors) with θ=+45°, θ=+90° and θ=+135°. Second, the accuracies of the NM models were low from the top (ϕ=0°), bottom (ϕ=180°), back (θ=−180°) and front (θ=0°) views which were then the inefficient views. Third, the efficient views, the views from where the accuracies of these models were high, were the sided-back (θ=±135°), side (θ=±90°), sided-front (θ=±45°), middle-top (ϕ =+45°), middle (ϕ =+90°) and middle-bottom (ϕ =+135°) views.

I computed a Pearson correlation coefficient to statistically assess how equivalent the view accuracies of the NM models of this study and the view accuracies of the recurrent models of chapter 5 were. Since there was only one datapoint for each view in this correlation, and since there were 40 views for both studies, then there was a total of 40 datapoints in this correlation. The correlation matrix in Figure 7.5 shows there was a strong positive correlation between the view accuracies of the passive NM models and the passive recurrent models, *r*(38) = .99, *p* < .001. The efficient and inefficient views were equal for both types of passive models. there are two explanations for this equivalence. One, the architectures of the two groups of models were almost the same. Two, both groups of models were trained, validated and tested with the same ten feature datasets and the same ten splits of the datasets into three groups of training, validation and test data.

7.4.4 Efficient Selection of the Active Models

The second objective of this study was to look at whether the active SCM and RSCM models selected more often the efficient views rather than the others. The percentage frequencies of the selected views by both active groups at the fifth timepoints of the second phase did not vary with respect to the view angles θ. Nevertheless, they were a lot different for the different view angles ϕ in both active models.

The most frequently selected views by the SCM models were the bottom views (ϕ=180°).

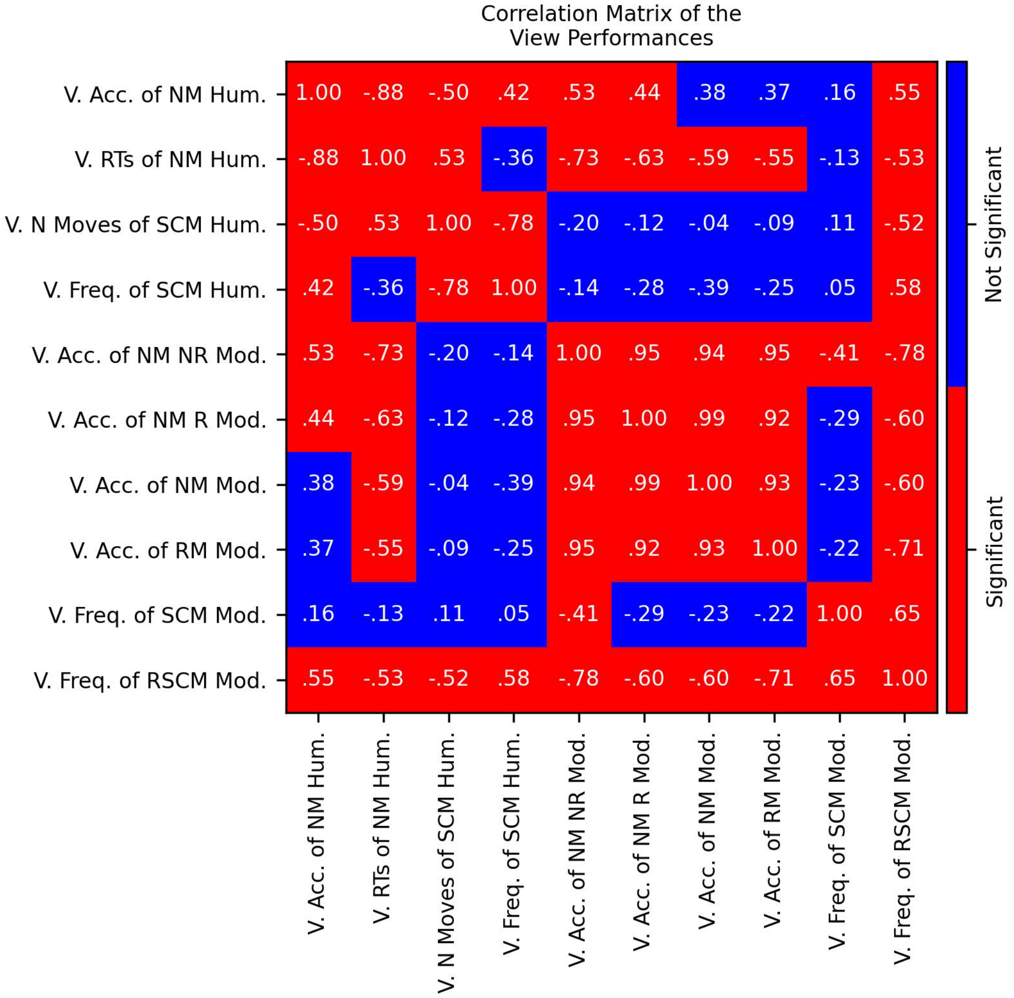


(A) View Percentage Frequency of the SCM Models at TimePoint 4 after the Training Phase 1

(B) View Percentage Frequency of the RSCM Models at TimePoint 4 after the Training Phase 1

Figure 7.4. The percentage frequncies of the selected views by the SCM (A) and the RSCM (B) groups at the fifth timepoint of the second experiment phase.

The top views (ϕ=0°) were the second most selected views by the SCM models. Then, the middle (ϕ=90°) and middle-bottom (ϕ=135°) views were slightly less selected. The least selected views by the SCM models were the middle-top views (ϕ=45°). The top views (ϕ=0°) were the most selected views by the RSCM models. The bottom views (ϕ=180°) were marginally less selected by the same models than the top ones. The third most



	V. Acc. of NM Hum.	V. RTs of NM Hum.	V. N Moves of SCM Hum.	V. Freq. of SCM Hum.	V. Acc. of NM NR Mod.	V. Acc. of NM R Mod.	V. Acc. of NM Mod.	V. Acc. of RM Mod.	V. Freq. of SCM Mod.	V. Freq. of RSCM Mod.
V. Acc. of NM Hum.	1.00	-.88	-.50	.42	.53	.44	.38	.37	.16	.55
V. RTs of NM Hum.	-.88	1.00	.53	-.36	-.73	-.63	-.59	-.55	-.13	-.53
V. N Moves of SCM Hum.	-.50	.53	1.00	-.78	-.20	-.12	-.04	-.09	.11	-.52
V. Freq. of SCM Hum.	.42	-.36	-.78	1.00	-.14	-.28	-.39	-.25	.05	.58
V. Acc. of NM NR Mod.	.53	-.73	-.20	-.14	1.00	.95	.94	.95	-.41	-.78
V. Acc. of NM R Mod.	.44	-.63	-.12	-.28	.95	1.00	.99	.92	-.29	-.60
V. Acc. of NM Mod.	.38	-.59	-.04	-.39	.94	.99	1.00	.93	-.23	-.60
V. Acc. of RM Mod.	.37	-.55	-.09	-.25	.95	.92	.93	1.00	-.22	-.71
V. Freq. of SCM Mod.	.16	-.13	.11	.05	-.41	-.29	-.23	-.22	1.00	.65
V. Freq. of RSCM Mod.	.55	-.53	-.52	.58	-.78	-.60	-.60	-.71	.65	1.00

Figure 7.5. The correlation matrix of some dependent variables of the views.

selected views were the middle views (ϕ=90°), while the least selected views were the middle-top (ϕ=45°) and the middle-bottom (ϕ=135°) views. The view selection of the SCM and the RSCM models did not match the pattern of efficient and inefficient views of the passive NM models.

Two Pearson correlation coefficients were computed to evaluate the linear relationships between the percentage frequencies of the selected views for both the active groups (Figure 7.4) and the action recognition accuracies of the views for the passive NM group (Figure 7.3). Both correlation coefficients are included in the correlation matrix of Figure 7.5. There was a marginal negative correlation between the selected view frequencies of

SCM models and the view accuracies of the NM models, *r*(38) = −.23, *p* < .159. There was a large negative correlation between the selected view frequencies of the RSCM models and the view accuracies of the NN models, *r*(38) = −.60, *p* < .001. These coefficients were not in line with my predictions since they were expected to be both significantly positive if active models selected the efficient view more often than the inefficient view.

The investigation of the linear relationships between the selected view frequencies of the active models and the passive recurrent models in chapter 5 was not needed, because the correlation coefficient between the view accuracies of the passive NM models in this chapter and the view accuracies of the passive recurrent models in chapter 5 (Figure 5.7) was almost 1 (.99). Thus, we can assume that the linear relationships between the selected view frequencies of the active models and the view accuracies of the passive recurrent models in chapter 5 were approximately the same as the relationships of the selected view frequencies of the active models and the view accuracies of the passive NM models in this chapter.

7.5 Conclusions

The main goal was this study was to investigate whether robotic observers can learn to process efficient active action classifications by DQN. Therefore, I trained passive and active models to classify the actions in dynamic videos by supervised learning and to select the next best views for more accurate action classifications by DQN. Both groups of active models classified the actions more accurately than both groups of passive models. The action recognition of the most accurate active group, the SCM group, was 1.5% more accurate than the most accurate passive group, the RM group. These results suggest that robotic observers can indeed learn efficient active action classification by DQN. These results were also in line with other studies about active computer vision (Arzati &

Arzanpour, 2021; Jayaraman & Grauman, 2018; Ramakrishnan & Grauman, 2018; Roost et al., 2020), where active models which had trained to select the next best view by DQN outperform the passive models in another task such as body pose estimation, visual perspective taking, and more.

However, the view selection of the active models was not positively correlated with the view efficiency of the passive NMs models. Thus, they did not select the efficient views more often than the inefficient ones. Yet, other results which were not included in this thesis indicated that the active models tended to obliquely move their viewpoints far away from their starting viewpoints. My explanation is that the action recognitions of models are more accurate if they watch the actions from many efficient and inefficient views rather than from only one single efficient view. Therefore, the active models may have learned to watch the same actions from many viewpoints at different timepoints by moving their viewpoints far and obliquely. In this way, the active models ultimately increased their action recognition accuracy.

8 Discussions

8.1 Conclusions

This thesis aimed to discover whether and how both human and robotic observers process efficient active action recognition. Thus, this thesis had seven objectives with that major aim. The first objective of the thesis was to choose an action dataset for my active action recognition studies. An action dataset can be suitable for active action recognition only if it includes multiple viewpoints from where the observers can watch the same actions. In this way, the observers can choose a viewpoint from multiple options and then the researcher can assess how efficient this choice was. The published action datasets are not appropriate for active action recognition either because of at least one of three issues. One, they only contain visual data (images or videos) that show the same actions from only one single viewpoint. Two, their visual data show the same actions from some very limited viewpoints up to three. Three, they are not 3D simulators that render images from the chosen viewpoints in 3D scenes with 3D acting actors. Thus, I made MVVHA which has MVVs showing the same actions from many viewpoints up to 40 in the most recent versions.

The second objective was to highlight the efficient and inefficient views for the action recognition of human observers. By looking at the action recognition accuracies and RTs of the human participants from different viewpoints with only NMs, I identified the efficient and inefficient views for action recognition of humans. The efficient views for the action recognition of human observers were the half-sided front (right-front and left-front) and the middle-height views, whereas the inefficient views were the back, the upper and the lower views. Additionally, the top views were slightly more efficient than the bottom views and the front views tended to be more efficient than the back views.

The third objective of the thesis was to identify the efficient and inefficient views for the action recognition of robotic observers. Thus, by computing the action recognition accuracies and classification losses of some basic computer models from different views without any viewpoint movements, I distinguished the efficient and inefficient views for the action recognition of robots. The efficient views for the computer models were the side (right and left), the half-sided (right-back, left-back, right-front and the left-front) and the middle-height views with respect to the actor. The inefficient views were the back, the front, the upper and the lower views. The bottom views were slightly more efficient than the top views.

Moreover, the pattern of efficient and inefficient views for action recognition was similar for both humans and robots. This conclusion was indicated by the fact that the view accuracies of the models with no view movements were correlated with the view accuracies and with view RTs of the human participants without any view movements. This suggests two further conclusions. One, the efficient and the inefficient views are independent of the observer type. Two, the NNs can approximate the complicated function of human vision, even if the NNs have many properties that are not biologically plausible.

Since the action recognition performance of the human and robotic observers was sensitive to the viewpoint, I encourage the scientific communities of computer vision and human vision to consider different viewpoints when elaborating vision theories and designing vision studies. For instance, I suggest controlling the viewpoints among the conditions of any visual study, by showing the same visual stimuli to the participants from different viewpoints. I additionally suggest verifying whether the results of a vision study are either the same or different from different viewpoints.

The fourth objective was to evaluate whether humans recognise the action in the active SCM condition than in the passive RM condition. Chapters 3 and 6 found small evidence supporting the advantage of human action recognition in the SCM condition over the RM condition and these were not easily replicated by different studies with different methods. The pilot study of Chapter 3 manipulated the two conditions of view movement type within-subject. The participants were tested in the lab with very clear images. The actions in the dynamic videos were perfectly controlled since they were identical for the two conditions. However, the two studies of chapter 6 manipulated the movement type between-subject. The participants were tested online without the pressure of proximity with an experimenter. The actions of the dynamic videos were randomly selected for each participant, and they were not perfectly controlled between the two conditions. There were two differences between the two studies of chapter 6. The first difference was that the video images were very unclear in the first experiment and slightly unclear in the second one. The second difference was that the trials started with the first frames of the videos in the first experiment, while they started with very clear T-poses.

The results of these studies were quite conflicting. In chapter 3, the human action recognition in the SCM condition was significantly faster than in the RM condition by 84 ms. Yet, the active action recognition of the SCM trials was slightly more accurate by 0.4% than the passive action recognition of the RM trials and this difference in accuracy was not significant. In the first study of chapter 6, the human participants in the SCM condition recognised the actions 3.5% more accurately and 167 ms more quickly than in the RM condition. Both the differences in accuracy and RT between the two conditions were not statistically significant. In the second study of chapter 6, these differences were tiny and neglectable.

My explanation of these controversial results is that the advantage of the SCM condition over the RM condition in the action recognition of humans interacts with the clearness of the images. The advantage is small with clear images and large with unclear images. If the images are clear, only the difference in RT is significant while the difference in accuracy is neglectable. On the other hand, if the images are very unclear, both the differences in accuracy and RT are significant. My view is in line with the results of the pilot study of chapter 3, but it is not consistent with the two studies of chapter 6 with almost clear images. In the first study of chapter 6, the differences in accuracy and RT were large and in the right direction, but the p-values were insignificant. The contrasting results of the first study of chapter 6 were that both the p-values of the differences in accuracy and RT were not significant, but these differences were large and in the right direction. The conflicting result of the second study of chapter 6 was that the difference in RT was very insignificant.

The results of the two studies of chapter 6 may have been incongruent with my view because of three limitations which may have added noise to the performances and reduced the statistical power. The most important limit was the manipulation of the movement type being between-subject. The second limit was the fact that the participants may have not done properly the task because they did it online without the pressure of any experimenter. The third limit was that the actions of the videos were not perfectly controlled across the two conditions and the actions recognised by the participants in one condition may have been easier than the actions in the other condition.

The fifth objective was to investigate whether human observers in the SCM condition select the efficient views more often than the inefficient views during active action recognition. The different studies unanimously showed humans selected the efficient viewpoints more frequently than the inefficient views. Chapter 3 showed that the different human frequencies of selected views in the SCM condition were positively correlated with human accuracies of the different starting views and negatively with human RTs of the different starting views. Because the views moved within trials, they were not well controlled between the different view conditions. To overcome this limitation, chapter 4 studied the view accuracies and the view RTs of humans in the different views with NMs. Then, both studies in chapter 6 showed that the frequencies of the selected views by the SCM participants were positively correlated with the view accuracies and negatively with the view RTs of the NM participants in chapter 4.

The sixth objective was to assess whether the active robots can recognise the actions more accurately than the corresponding passive robots. In the study of chapter 7, all passive and active models were trained to classify the actions by supervised learning and to select the next best view for more accurate action recognition by DQN. Both groups of active models, the SCM and the RSCM groups, were overall more accurate than both the passive models, the RM and the NM groups. The most accurate active group, the SCM group, was 1.5% more accurate than the most accurate passive group, the RM group.

The seventh and final objective was to investigate whether the active models select the efficient views more than the inefficient views when they actively classify the actions. The active SCM and RSCM models of chapter 7 did not choose the efficient views more frequently than the inefficient ones, since their frequencies of selected views were not correlated with the view accuracies of the NM models of both studies of chapters 5 and 7. Instead, these active models tended to move their viewpoints obliquely to viewpoints which were far from the starting view. My best interpretation of this efficient view selection of the active models is that the active models may have learned through the DQN method that they can achieve more accurate action classification by observing the same actions from multiple viewpoints selected at different timepoints rather than from only a few efficient viewpoints.

8.2 Biological Plausibility Limits

From the perspective of computational neuroscience, there were at least five limits in the biological plausibility of the RCNNs in this thesis. One, the neural activation function was ReLU (Nair & Hinton, 2010), mostly in the ResNets layers, while the activation of biological neurons is binary or all-or-none (Hodgkin & Huxley, 1990). Two, their numbers of layers were not biologically plausible. The models with 18 to 20 layers are a lot deeper than the visual cortical areas (Liao & Poggio, 2016) which are involved in action recognition. Three, the models only have one recurrent layer (LSTM) with a lateral connection. The other layers only have forward connections and do not have any lateral or backward connections. However, the visual cortical areas are FRCNNs (Liao & Poggio, 2016) given that they have forward, lateral and backward connections (Kar et al., 2019; Kubilius et al., 2018; Lamme et al., 1998). Four, despite some forward shortcuts being in the ResNets of my models, there are no backward shortcut connections in my models. There are both forward and backward shortcuts connecting the visual areas of the cortex. Each visual area sends output information to all visual areas, including itself, and receives input information from all visual areas, including itself. Five, they do not simulate action recognition RTs of the brain. It may appear they have RTs in the range of 1.33 and 1.67 seconds given that they show the highest accuracy at the fourth and fifth frames of 3-Hz videos. This only shows that the actions are objectively more evident for the robotic observers in the (fourth and fifth) frames which are within the time interval from 1.33 to 1.67 of the 2-second videos. However, the computational models do not tell anything about how long it takes for the brain to process and classify actions after the stimulus onsets of the fourth and fifth frames.

8.3 Future Research

The future research in robotic active action recognition must tackle these issues concerning the biological implausibility of the RCNNs. I suggest doing the next experiments with FRCNNs which are more biological plausible. The layers must be 4 to 8 fully-recurrent layers and their activation refresh rate of the layers must be between 20hz and 50hz as suggested by Liao and Poggio (2016). Additionally, sigmoid and ideally all-or-none activation functions must be implemented because they are more biological plausible.

The activation of any fully recurrent layer at any timepoint t is a non-linear function of the activations of all layers, including itself, at the previous timepoint t−1. Each direct connection between two layers of a fully recurrent model has its own parameters (weights and biases) which can be optimized. Therefore, the FRCNNs have forward backward and lateral connection. The forward and backward connections can either be shortcut or simple connections. These models may accurately predict the human RTs of active action recognition.

Future psychology experiments must investigate whether only within-subject designs can show statistically significant improvements in the action recognition accuracy and RT of humans with active vision rather than passive vision. If that is the case, this would explain why my between-subject online studies did not show significant differences in the performances of active and passive conditions.

Further research in human active action recognition would improve our knowledge of the underlining mechanisms of the human brain involved in active action recognition. That will stimulate the innovation of more intelligent and more biological active vision models for active action recognition. Thus, the future robots will be more social with people, by being able to read their non-verbal language. Moreover, automatic security surveillance and health monitoring will be more accurate and prevent crimes and illnesses.

9 References

Adams, R. B., & Janata, P. (2002). A comparison of neural circuits underlying auditory and visual object categorization [Article]. *Neuroimage*, *16*(2), 361-377. https://doi.org/10.1006/nimg.2002.1088

Aghaei, A., Nazari, A., & Moghaddam, M. E. (2021). Sparse deep LSTMs with convolutional attention for human action recognition. *SN Computer Science*, *2*(3), 1-14.

Al-Faris, M., Chiverton, J., Ndzi, D., & Ahmed, A. I. (2020). A Review on Computer Vision-Based Methods for Human Action Recognition [Review]. *Journal of Imaging*, *6*(6), 32, Article 46. https://doi.org/10.3390/jimaging6060046

Al-Faris, M., Chiverton, J. P., Yang, Y. U. Y., & Ndzi, D. L. (2020). Multi-view region-adaptive multi-temporal DMM and RGB action recognition [Article]. *Pattern Analysis and Applications*, *23*(4), 1587-1602. https://doi.org/10.1007/s10044-020-00886-5

Amalric, M., & Dehaene, S. (2016). Origins of the brain networks for advanced mathematics in expert mathematicians [Article]. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(18), 4909-4917. https://doi.org/10.1073/pnas.1603205113

Amalric, M., & Dehaene, S. (2018). Cortical circuits for mathematical knowledge: evidence for a major subdivision within the brain's semantic networks [Review]. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *373*(1740), 9, Article 20160515. https://doi.org/10.1098/rstb.2016.0515

Amalric, M., & Dehaene, S. (2019). A distinct cortical network for mathematical knowledge in the human brain [Article]. *Neuroimage*, *189*, 19-31. https://doi.org/10.1016/j.neuroimage.2019.01.001

Arzati, M. A., & Arzanpour, S. (2021, Oct 12-15). Viewpoint Selection for DermDrone using Deep Reinforcement Learning.*International Conference on Control Automation and Systems* [2021 21st international conference on control, automation and systems (iccas 2021)]. 21st International Conference on Control, Automation and Systems (ICCAS), South Korea.

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions [Article]. *Current Biology*, *16*(18), 1818-1823. https://doi.org/10.1016/j.cub.2006.07.060

Bart, E., & Hegdé, J. (2012). Invariant Recognition of Visual Objects: Some Emerging Computational Principles [Editorial]. *Frontiers in Computational Neuroscience*, *6*. https://doi.org/10.3389/fncom.2012.00060

Basu, M. (2002). Gaussian-based edge-detection methods - A survey [Article]. *Ieee Transactions on Systems Man and Cybernetics Part C-Applications and Reviews*, *32*(3), 252-260. https://doi.org/10.1109/tsmcc.2002.804448

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America a-Optics Image Science and Vision*, *20*(7), 1391-1397. https://doi.org/10.1364/josaa.20.001391

Bengio, Y., Simard, P., & Frasconi, P. (1994). LEARNING LONG-TERM DEPENDENCIES WITH GRADIENT DESCENT IS DIFFICULT [Article]. *Ieee Transactions on Neural Networks*, *5*(2), 157-166. https://doi.org/10.1109/72.279181

Biederman, I. (1987). RECOGNITION-BY-COMPONENTS - A THEORY OF HUMAN IMAGE UNDERSTANDING [Article]. *Psychological Review*, *94*(2), 115-147. https://doi.org/10.1037/0033-295x.94.2.115

Biederman, I., & Bar, M. (1999). One-shot viewpoint invariance in matching novel objects [Article]. *Vision Research*, *39*(17), 2885-2899. https://doi.org/10.1016/s0042-6989(98)00309-5

Biederman, I., & Bar, M. (2000). Differing views on views: response to Hayward and Tarr (2000) [Article]. *Vision Research*, *40*(28), 3901-3905. https://doi.org/10.1016/s0042-6989(00)00180-2

Biederman, I., & Gerhardstein, P. C. (1993). RECOGNIZING DEPTH-ROTATED OBJECTS - EVIDENCE AND CONDITIONS FOR 3-DIMENSIONAL VIEWPOINT INVARIANCE. *Journal of Experimental Psychology-Human Perception and Performance*, *19*(6), 1162-1182. https://doi.org/10.1037/0096-1523.19.6.1162

Biederman, I., & Gerhardstein, P. C. (1995). VIEWPOINT-DEPENDENT MECHANISMS IN VISUAL OBJECT RECOGNITION - REPLY TO TARR AND BULTHOFF (1995) [Article]. *Journal of Experimental Psychology-Human Perception and Performance*, *21*(6), 1506-1514. https://doi.org/10.1037/0096-1523.21.6.1506

Binder, J. R. (2015). The Wernicke area Modern evidence and a reinterpretation [Review]. *Neurology*, *85*(24), 2170-2175. https://doi.org/10.1212/wnl.0000000000002219

Blasing, B. E., & Sauzet, O. (2018). My Action, My Self: Recognition of Self-Created but Visually Unfamiliar Dance-Like Actions From Point-Light Displays [Article]. *Frontiers in Psychology*, *9*, 9, Article 1909. https://doi.org/10.3389/fpsyg.2018.01909

Boussaoud, D., Ungerleider, L. G., & Desimone, R. (1990). PATHWAYS FOR MOTION ANALYSIS - CORTICAL CONNECTIONS OF THE MEDIAL SUPERIOR TEMPORAL AND FUNDUS OF THE SUPERIOR TEMPORAL VISUAL AREAS IN THE MACAQUE [Article]. *Journal of Comparative Neurology*, *296*(3), 462-495. https://doi.org/10.1002/cne.902960311

Bradski, G. (2000). The OpenCV library [Article]. *Dr Dobbs Journal*, *25*(11), 120-123.

Bricolo, E., Poggio, T., & Logothetis, N. (1997). 3D object recognition: A model of view-tuned neurons [Proceedings Paper]. *Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference*, *9*, 41-47.

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., . . . Freund, H. J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study [Article]. *European Journal of Neuroscience*, *13*(2), 400-404. https://doi.org/10.1046/j.1460-9568.2001.01385.x

Buccino, G., Binkofski, F., & Riggio, L. (2004). The mirror neuron system and action recognition. *Brain and Language*, *89*(2), 370-376. https://doi.org/10.1016/s0093-934x(03)00356-0

Chaaraoui, A. A., Climent-Perez, P., & Florez-Revuelta, F. (2012). A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living [Review]. *Expert Systems with Applications*, *39*(12), 10873-10888. https://doi.org/10.1016/j.eswa.2012.03.005

Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., & Deng, J. (2018). Learning to detect human-object interactions. 2018 ieee winter conference on applications of computer vision (wacv),

Chao, Y.-W., Wang, Z., He, Y., Wang, J., & Deng, J. (2015). Hico: A benchmark for recognizing human-object interactions in images. Proceedings of the IEEE International Conference on Computer Vision,

Chong, T. T. J., Cunnington, R., Williams, M. A., Kanwisher, N., & Mattingley, J. B. (2008). fMRI Adaptation Reveals Mirror Neurons in Human Inferior Parietal Cortex. *Current Biology*, *18*(20), 1576-1580. https://doi.org/10.1016/j.cub.2008.08.068

Cooper, E. E., & Wojan, T. J. (2000). Differences in the coding of spatial relations in face identification and basic-level object recognition [Article; Proceedings Paper]. *Journal of Experimental Psychology-Learning Memory and Cognition*, *26*(2), 470-488.

https://doi.org/10.1037//0278-7393.26.2.470

Corballis, M. C., Zbrodoff, N. J., Shetzer, L. I., & Butler, P. B. (1978). DECISIONS ABOUT IDENTITY AND ORIENTATION OF ROTATED LETTERS AND DIGITS [Article]. *Memory & Cognition*, *6*(2), 98-107. https://doi.org/10.3758/bf03197434

Dai, C., Liu, X., & Lai, J. (2020). Human action recognition using two-stream attention based LSTM networks. *Applied soft computing*, *86*, 105820.

Daugman, J. G. (1985). UNCERTAINTY RELATION FOR RESOLUTION IN SPACE, SPATIAL-FREQUENCY, AND ORIENTATION OPTIMIZED BY TWO-DIMENSIONAL VISUAL CORTICAL FILTERS [Article]. *Journal of the Optical Society of America a-Optics Image Science and Vision*, *2*(7), 1160-1169. https://doi.org/10.1364/josaa.2.001160

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Li, F. F., & Ieee. (2009, Jun 20-25). ImageNet: A Large-Scale Hierarchical Image Database.*IEEE Conference on Computer Vision and Pattern Recognition* [Cvpr: 2009 ieee conference on computer vision and pattern recognition, vols 1-4]. IEEE-Computer-Society Conference on Computer Vision and Pattern Recognition Workshops, Miami Beach, FL.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015, Jun 07-12). Long-term Recurrent Convolutional Networks for Visual Recognition and Description.*IEEE Conference on Computer Vision and Pattern Recognition* [2015 ieee conference on computer vision and pattern recognition (cvpr)]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA.

Du, S., Wang, Y., Zhai, X., Balakrishnan, S., Salakhutdinov, R., & Singh, A. (2018). How many samples are needed to learn a convolutional neural network? *Stat*, *1050*, 21.

Du, S., Wang, Y., Zhai, X., Balakrishnan, S., Salakhutdinov, R., & Singh, A. (2019). How many samples are needed to estimate a convolutional or recurrent neural network? *Stat*, *1050*, 30.

Edelman, S., & Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations [Review]. *Cognitive Science*, *27*(1), 73-109, Article Pii s0364-0213(02)00114-3.

Fairhall, S. L., & Caramazza, A. (2013). Brain Regions That Represent Amodal Conceptual Knowledge [Article]. *Journal of Neuroscience*, *33*(25), 10552-10558. https://doi.org/10.1523/jneurosci.0051-13.2013

Ferstl, Y., Bulthoff, H., & de la Rosa, S. (2017). Action recognition is sensitive to the identity of the actor [Article]. *Cognition*, *166*, 201-206. https://doi.org/10.1016/j.cognition.2017.05.036

Foster, D. H., & Gilson, S. J. (2002). Recognizing novel three-dimensional objects by summing signals from parts and views [Article]. *Proceedings of the Royal Society B-Biological Sciences*, *269*(1503), 1939-1947. https://doi.org/10.1098/rspb.2002.2119

Foulsham, T., & Lock, M. (2015). How the eyes tell lies: Social gaze during a preference task. *Cognitive science*, *39*(7), 1704-1726.

Foxe, J. J., Wylie, G. R., Martinez, A., Schroeder, C. E., Javitt, D. C., Guilfoyle, D., . . . Murray, M. M. (2002). Auditory-somatosensory multisensory processing in auditory association cortex: An fMRI study [Article]. *Journal of Neurophysiology*, *88*(1), 540-543. https://doi.org/10.1152/jn.2002.88.1.540

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325-1352. https://doi.org/10.1016/j.neunet.2003.06.005

Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1-3), 70-87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Gabor, D. (1946). Theory of communication. *Journal of the Institution of Electrical Engineers-part III: radio and communication engineering*, *93*(26), 429-441. https://doi.org/10.1049/ji-3-2.1946.0074

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, *119*, 593-609. https://doi.org/10.1093/brain/119.2.593

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (2002). Action representation and the inferior parietal lobule [Article; Proceedings Paper]. *Common Mechanisms in Perception and Action*, *19*, 334-355.

Galletti, C., Kutz, D. F., Gamberini, M., Breveglieri, R., & Fattori, P. (2003). Role of the medial parieto-occipital cortex in the control of reaching and grasping movements [Article; Proceedings Paper]. *Experimental Brain Research*, *153*(2), 158-170. https://doi.org/10.1007/s00221-003-1589-z

Girshick, R. (2015, Dec 11-18). Fast R-CNN.*IEEE International Conference on Computer Vision* [2015 ieee international conference on computer vision (iccv)]. IEEE International Conference on Computer Vision, Santiago, CHILE.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics,

Goodale, M. A., & Milner, A. D. (1992). SEPARATE VISUAL PATHWAYS FOR PERCEPTION AND ACTION [Article]. *Trends in Neurosciences*, *15*(1), 20-25. https://doi.org/10.1016/0166-2236(92)90344-8

Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., . . . Sukthankar, R. (2018). Ava: A video dataset of spatio-temporally localized atomic visual actions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Hafri, A., Trueswell, J. C., & Epstein, R. A. (2017). Neural Representations of Observed

Actions Generalize across Static and Dynamic Visual Input [Article]. *Journal of Neuroscience*, *37*(11), 3056-3071. https://doi.org/10.1523/jneurosci.2496-16.2017

Hamm, J. P., & McMullen, P. A. (1998). Effects of orientation on the identification of rotated objects depend on the level of identity [Article]. *Journal of Experimental Psychology-Human Perception and Performance*, *24*(2), 413-426. https://doi.org/10.1037/0096-1523.24.2.413

Hamzei, F., Rijntjes, M., Dettmers, C., Glauche, V., Weiller, C., & Buchel, C. (2003). The human action recognition system and its relationship to Broca's area: an fMRI study [Article]. *Neuroimage*, *19*(3), 637-644. https://doi.org/10.1016/s1053-8119(03)00087-9

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., . . . Oliphant, T. E. (2020). Array programming with NumPy [Review]. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Hayward, W. G., & Tarr, M. J. (1997). Testing conditions for viewpoint invariance in object recognition [Article]. *Journal of Experimental Psychology-Human Perception and Performance*, *23*(5), 1511-1521. https://doi.org/10.1037/0096-1523.23.5.1511

Hayward, W. G., & Tarr, M. J. (2000). Differing views on views: comments on Biederman and Bar (1999) [Article]. *Vision Research*, *40*(28), 3895-3899. https://doi.org/10.1016/s0042-6989(00)00179-6

Hayward, W. G., & Williams, P. (2000). Viewpoint dependence and object discriminability [Article]. *Psychological Science*, *11*(1), 7-12. https://doi.org/10.1111/1467-9280.00207

He, K. M., Zhang, X. Y., Ren, S. Q., & Sun, J. (2016, Jun 27-30). Deep Residual Learning for Image Recognition.*IEEE Conference on Computer Vision and Pattern Recognition* [2016 ieee conference on computer vision and pattern recognition (cvpr)]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA.

He, K. M., Zhang, X. Y., Ren, S. Q., Sun, J., & Ieee. (2015, Dec 11-18). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.*IEEE International Conference on Computer Vision* [2015 ieee international conference on computer vision (iccv)]. IEEE International Conference on Computer Vision, Santiago, CHILE.

He, K. M., Zhang, X. Y., Ren, S. Q., Sun, J., & Ieee. (2016, Jun 27-30). Deep Residual Learning for Image Recognition.*IEEE Conference on Computer Vision and Pattern Recognition* [2016 ieee conference on computer vision and pattern recognition (cvpr)]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA.

Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey [Article]. *Image and Vision Computing*, *60*, 4-21. https://doi.org/10.1016/j.imavis.2017.01.010

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory [Article]. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hodgkin, A. L., & Huxley, A. F. (1990). A QUANTITATIVE DESCRIPTION OF MEMBRANE CURRENT AND ITS APPLICATION TO CONDUCTION AND EXCITATION IN NERVE (REPRINTED FROM JOURNAL OF PHYSIOLOGY, VOL 117, PG 500-544, 1952) [Article; Proceedings Paper]. *Bulletin of Mathematical Biology*, *52*(1-2), 25-71. https://doi.org/10.1007/bf02459568

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision:

Implications for shape perception and object recognition [Article]. *Visual Cognition*, *8*(3-5), 489-517. https://doi.org/10.1080/13506280143000214

Hummel, J. E., & Biederman, I. (1992). DYNAMIC BINDING IN A NEURAL NETWORK FOR SHAPE-RECOGNITION [Article]. *Psychological Review*, *99*(3), 480-517. https://doi.org/10.1037/0033-295x.99.3.480

Hummel, J. E., & Stankiewicz, B. J. (1996). Categorical relations in shape perception [Article]. *Spatial Vision*, *10*(3), 201-236. https://doi.org/10.1163/156856896x00141

Ioffe, S., & Szegedy, C. (2015, Jul 07-09). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.*Proceedings of Machine Learning Research* [International conference on machine learning, vol 37]. 32nd International Conference on Machine Learning, Lille, FRANCE.

Ishai, A., Ungerleider, L. G., & Haxby, J. V. (2000). Distributed neural systems for the generation of visual images [Article]. *Neuron*, *28*(3), 979-990. https://doi.org/10.1016/s0896-6273(00)00168-9

Jayaraman, D., & Grauman, K. (2018). Learning to look around: intelligently exploring unseen environments for unknown tasks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT.

Ji, S. W., Xu, W., Yang, M., & Yu, K. (2013). 3D Convolutional Neural Networks for Human Action Recognition [Article]. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221-231. https://doi.org/10.1109/tpami.2012.59

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior [Article]. *Nature Neuroscience*, *22*(6), 974-+. https://doi.org/10.1038/s41593-019-0392-5

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., & Ieee. (2014, Jun 23-28). Large-scale Video Classification with Convolutional Neural Networks.*IEEE Conference on Computer Vision and Pattern Recognition* [2014 ieee conference on computer vision and pattern recognition (cvpr)]. 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH.

Kennedy, L. M., & Basu, M. (1997). Image enhancement using a human visual system model [Article]. *Pattern Recognition*, *30*(12), 2001-2014. https://doi.org/10.1016/s0031-3203(97)00014-9

Keysers, C., Kohler, E., Umilta, M. A., Nanetti, L., Fogassi, L., & Gallese, V. (2003). Audiovisual mirror neurons and action recognition [Article; Proceedings Paper]. *Experimental Brain Research*, *153*(4), 628-636. https://doi.org/10.1007/s00221-003-1603-5

Knoblich, G., & Sebanz, N. (2006). The social nature of perception and action [Article]. *Current Directions in Psychological Science*, *15*(3), 99-104. https://doi.org/10.1111/j.0963-7214.2006.00415.x

Kording, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244-247. https://doi.org/10.1038/nature02169

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping [Article]. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863-3868. https://doi.org/10.1073/pnas.0600244103

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet Classification with Deep Convolutional Neural Networks [Article]. *Communications of the Acm*, *60*(6), 84-90. https://doi.org/10.1145/3065386

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*, 408385. https://doi.org/10.1101/408385

Kuffler, S. W. (1952). Neurons in the retina: organization, inhibition and excitation problems. Cold Spring Harbor Symposia on Quantitative Biology,

Lamme, V. A. F., Super, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex [Review]. *Current Opinion in Neurobiology*, *8*(4), 529-535. https://doi.org/10.1016/s0959-4388(98)80042-1

Lawson, R., & Humphreys, G. W. (1998). View-specific effects of depth rotation and foreshortening on the initial recognition and priming of familiar objects [Article]. *Perception & Psychophysics*, *60*(6), 1052-1066. https://doi.org/10.3758/bf03211939

Le, D.-T., Uijlings, J., & Bernardi, R. (2014). Tuhoi: Trento universal human object interaction dataset. Proceedings of the Third Workshop on Vision and Language,

LeCun, Y., Bottou, L., Orr, G. B., & Muller, K. R. (1998). Efficient backprop. *Neural Networks: Tricks of the Trade*, *1524*, 9-50. https://doi.org/10.1007/3-540-49430-8_2

Lee, C. Y., Xie, S. N., Gallagher, P. W., Zhang, Z. Y., & Tu, Z. W. (2015, May 09-12). Deeply-Supervised Nets.*JMLR Workshop and Conference Proceedings* [Artificial intelligence and statistics, vol 38]. 18th International Conference on Artificial Intelligence and Statistics (AISTATS), San Diego, CA.

Lee, K., Ognibene, D., Chang, H. J., Kim, T. K., & Demiris, Y. (2015). STARE: Spatio-Temporal Attention Relocation for Multiple Structured Activities Detection [Article]. *Ieee Transactions on Image Processing*, *24*(12), 12. https://doi.org/10.1109/tip.2015.2487837

Leek, E. C., & Johnston, S. J. (2006). A polarity effect in misoriented object recognition: The role of polar features in the computation of orientation-invariant shape representations [Article]. *Visual Cognition*, *13*(5), 573-600. https://doi.org/10.1080/13506280544000048

Li, D., Qiu, Z. F., Dai, Q., Yao, T., & Mei, T. (2018). Recurrent Tubelet Proposal and Recognition Networks for Action Detection [Proceedings Paper]. *Computer Vision - Eccv 2018, Pt Vi*, *11210*, 306-322. https://doi.org/10.1007/978-3-030-01231-1_19

Liao, Q., & Poggio, T. (2016). Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. *arXiv*. https://doi.org/10.48550/ARXIV.1604.03640

Lingnau, A., & Downing, P. E. (2015). The lateral occipitotemporal cortex in action [Review]. *Trends in Cognitive Sciences*, *19*(5), 268-277. https://doi.org/10.1016/j.tics.2015.03.006

Liu, A.-A., Xu, N., Nie, W.-Z., Su, Y.-T., Wong, Y., & Kankanhalli, M. (2016). Benchmarking a multimodal and multiview and interactive dataset for human action recognition. *IEEE Transactions on cybernetics*, *47*(7), 1781-1794.

Liu, J., Shahroudy, A., Perez, M. L., Wang, G., Duan, L.-Y., & Chichung, A. K. (2019). Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*.

Liu, Y., Nie, L., Han, L., Zhang, L., & Rosenblum, D. S. (2015). Action2Activity: recognizing complex activities from sensor data. Twenty-fourth international joint conference on artificial intelligence,

Liu, Y., Nie, L., Liu, L., & Rosenblum, D. S. (2016). From action to activity: sensor-based activity recognition. *Neurocomputing*, *181*, 108-115.

Logothetis, N. K., Pauls, J., Bulthoff, H. H., & Poggio, T. (1994). VIEW-DEPENDENT OBJECT RECOGNITION BY MONKEYS [Article]. *Current Biology*, *4*(5), 401-414. https://doi.org/10.1016/s0960-9822(00)00089-0

Ma, S., Bargal, S. A., Zhang, J., Sigal, L., & Sclaroff, S. (2017). Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern*

Recognition, *68*, 334-345.

Majd, M., & Safabakhsh, R. (2020). Correlational Convolutional LSTM for human action recognition [Article]. *Neurocomputing*, *396*, 224-229. https://doi.org/10.1016/j.neucom.2018.10.095

Malach, R., Levy, I., & Hasson, U. (2002). The topography of high-order human object areas [Review]. *Trends in Cognitive Sciences*, *6*(4), 176-184. https://doi.org/10.1016/s1364-6613(02)01870-3

Man, K., Damasio, A., Meyer, K., & Kaplan, J. T. (2015). Convergent and invariant object representations for sight, sound, and touch [Article]. *Human Brain Mapping*, *36*(9), 3629-3640. https://doi.org/10.1002/hbm.22867

Mao, A., Mohri, M., & Zhong, Y. (2023). Cross-entropy loss functions: Theoretical analysis and applications. *arXiv preprint arXiv:2304.07288*.

Marcelja, S. (1980). MATHEMATICAL-DESCRIPTION OF THE RESPONSES OF SIMPLE CORTICAL-CELLS [Article]. *Journal of the Optical Society of America*, *70*(11), 1297-1300. https://doi.org/10.1364/josa.70.001297

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman and Company.

Marr, D., & Hildreth, E. (1980). THEORY OF EDGE-DETECTION [Article]. *Proceedings of the Royal Society Series B-Biological Sciences*, *207*(1167), 187-217. https://doi.org/10.1098/rspb.1980.0020

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *200*(1140), 269-294.

Meijer, D., Veselič, S., Calafiore, C., & Noppeney, U. (2019). Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation. *Cortex*, *119*, 74-88.

Milivojevic, B. (2012). Object recognition can be viewpoint dependent or invariant - it's just a matter of time and task [Editorial Material]. *Frontiers in Computational Neuroscience*, *6*, 3, Article 27. https://doi.org/10.3389/fncom.2012.00027

Milner, A. D., & Goodale, M. A. (2008). Two visual systems re-viewed [Article]. *Neuropsychologia*, *46*(3), 774-785. https://doi.org/10.1016/j.neuropsychologia.2007.10.005

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways [Review]. *Trends in Neurosciences*, *6*(10), 414-417. https://doi.org/10.1016/0166-2236(83)90190-X

Mitchell, R. W., & Curry, C. (2016). Self-Recognition and Other-Recognition in Point-Light Displays. *Open Journal of Philosophy*, *6*(01), 42-50. https://doi.org/10.4236/ojpp.2016.61005

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., . . . Vondrick, C. (2019). Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, *42*(2), 502-508.

Monfort, M., Ramakrishnan, K., Andonian, A., McNamara, B. A., Lascelles, A., Pan, B., . . . Oliva, A. (2019). Multi-Moments in Time: Learning and Interpreting Models for Multi-Action Video Understanding. *arXiv preprint arXiv:1911.00232*.

Montufar, G., Pascanu, R., Cho, K., & Bengio, Y. (2014). On the Number of Linear Regions of Deep Neural Networks [Proceedings Paper]. *Advances in Neural Information Processing Systems 27 (Nips 2014)*, *27*, 9.

Moors, A. (2016). Automaticity: Componential, Causal, and Mechanistic Explanations. In S. T. Fiske (Ed.), *Annual Review of Psychology, Vol 67* (Vol. 67, pp. 263-287).

Annual Reviews. https://doi.org/10.1146/annurev-psych-122414-033550

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis [Review]. *Psychological Bulletin*, *132*(2), 297-326. https://doi.org/10.1037/0033-2909.132.2.297

Moscovici, S. (2001). *Social Representations*. New York University Press.

Murata, A., Wen, W., & Asama, H. (2016). The body and objects represented in the ventral stream of the parieto-premotor network [Review]. *Neuroscience Research*, *104*, 4-15. https://doi.org/10.1016/j.neures.2015.10.010

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. Icml,

Ng, J. Y. H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015, Jun 07-12). Beyond Short Snippets: Deep Networks for Video Classification.*IEEE Conference on Computer Vision and Pattern Recognition* [2015 ieee conference on computer vision and pattern recognition (cvpr)]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA.

O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions [Article]. *Journal of Cognitive Neuroscience*, *12*(6), 1013-1023. https://doi.org/10.1162/08989290051137549

Oosterhof, N. N., Tipper, S. P., & Downing, P. E. (2013). Crossmodal and action-specific: neuroimaging the human mirror neuron system [Review]. *Trends in Cognitive Sciences*, *17*(7), 311-318. https://doi.org/10.1016/j.tics.2013.04.012

Orban, G. A., Lanzilotto, M., & Bonini, L. (2021). From Observed Action Identity to Social Affordances. *Trends in Cognitive Sciences*, *25*(6), 493-505. https://doi.org/10.1016/j.tics.2021.02.012

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Chintala, S.

(2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library [Proceedings Paper]. *Advances in Neural Information Processing Systems 32 (Nips 2019)*, *32*, 12.

Pietrini, P., Furey, M. L., Ricciardi, E., Gobbini, M. I., Wu, W. H. C., Cohen, L., . . . Haxby, J. V. (2004). Beyond sensory images: Object-based representation in the human ventral pathway [Article]. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(15), 5658-5663. https://doi.org/10.1073/pnas.0400707101

Poggio, T., & Edelman, S. (1990). A NETWORK THAT LEARNS TO RECOGNIZE 3-DIMENSIONAL OBJECTS [Article]. *Nature*, *343*(6255), 263-266. https://doi.org/10.1038/343263a0

Rai, M., & Rivas, P. (2020). A review of convolutional neural networks and gabor filters in object recognition. 2020 International Conference on Computational Science and Computational Intelligence (CSCI),

Ramakrishnan, S. K., & Grauman, K. (2018). Sidekick Policy Learning for Active Visual Exploration [Proceedings Paper]. *Computer Vision - Eccv 2018, Pt Xii*, *11216*, 424-442. https://doi.org/10.1007/978-3-030-01258-8_26

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects [Article]. *Nature Neuroscience*, *2*(1), 79-87. https://doi.org/10.1038/4580

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., & Ieee. (2016, Jun 27-30). You Only Look Once: Unified, Real-Time Object Detection.*IEEE Conference on Computer Vision and Pattern Recognition* [2016 ieee conference on computer vision and pattern recognition (cvpr)]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA.

Rizvi, S. T. H., Cabodi, G., Gusmao, P., & Francini, G. (2016). Gabor Filter based Image Representation for Object Classification [Proceedings Paper]. *2016 International Conference on Control, Decision and Information Technologies (Codit)*, 628-632.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, *3*(2), 131-141. https://doi.org/10.1016/0926-6410(95)00038-0

Rizzolatti, G., & Matelli, M. (2003). Two different streams form the dorsal visual system: anatomy and functions [Article; Proceedings Paper]. *Experimental Brain Research*, *153*(2), 146-157. https://doi.org/10.1007/s00221-003-1588-0

Rock, I., & Divita, J. (1987). A CASE OF VIEWER-CENTERED OBJECT PERCEPTION [Article]. *Cognitive Psychology*, *19*(2), 280-293. https://doi.org/10.1016/0010-0285(87)90013-2

Roost, D., Meier, R., Toffetti Carughi, G., & Stadelmann, T. (2020). *Combining reinforcement learning with supervised deep learning for neural active scene understanding* Active Vision and Perception in Human (-Robot) Collaboration Workshop at IEEE RO-MAN 2020 (AVHRC'20), online, 31 August-4 September 2020,

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyesbraem, P. (1976). BASIC OBJECTS IN NATURAL CATEGORIES [Article]. *Cognitive Psychology*, *8*(3), 382-439. https://doi.org/10.1016/0010-0285(76)90013-x

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge [Article]. *International Journal of Computer Vision*, *115*(3), 211-252. https://doi.org/10.1007/s11263-015-0816-y

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv*.

https://doi.org/10.48550/ARXIV.1402.1128

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2015). Functional lateralization of temporoparietal junction - imitation inhibition, visual perspective-taking and theory of mind [Article]. *European Journal of Neuroscience*, *42*(8), 2527-2533. https://doi.org/10.1111/ejn.13036

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Schraudolph, N. N. (1998). Centering neural network gradient factors [Article]. *Neural Networks: Tricks of the Trade*, *1524*, 207-226.

Schurz, M., Aichhorn, M., Martin, A., & Perner, J. (2013). Common brain areas engaged in false belief reasoning and visual perspective taking: a meta-analysis of functional brain imaging studies [Article]. *Frontiers in Human Neuroscience*, *7*, 14, Article 712. https://doi.org/10.3389/fnhum.2013.00712

Schutz-Bosbach, S., Mancini, B., Aglioti, S. M., & Haggard, P. (2006). Self and other in the human motor system [Article]. *Current Biology*, *16*(18), 1830-1834. https://doi.org/10.1016/j.cub.2006.07.048

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together [Review]. *Trends in Cognitive Sciences*, *10*(2), 70-76. https://doi.org/10.1016/j.tics.2005.12.009

Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. Proceedings of the IEEE conference on computer vision and pattern recognition,

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Spunt, R. P., & Lieberman, M. D. (2012). Dissociating Modality-Specific and Supramodal

Neural Systems for Action Understanding [Article]. *Journal of Neuroscience*, *32*(10), 3575-3583. https://doi.org/10.1523/jneurosci.5715-11.2012

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training Very Deep Networks [Proceedings Paper]. *Advances in Neural Information Processing Systems 28 (Nips 2015)*, *28*, 9.

Stankiewicz, B. J., Hummel, J. E., & Cooper, E. E. (1998). The role of attention in priming for left-right reflections of object images: Evidence for a dual representation of object shape [Article]. *Journal of Experimental Psychology-Human Perception and Performance*, *24*(3), 732-744. https://doi.org/10.1037/0096-1523.24.3.732

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting [Article]. *Journal of Cognitive Neuroscience*, *33*(10), 2044-2064. https://doi.org/10.1162/jocn_a_01755

Szegedy, C., Liu, W., Jia, Y. Q., Sermanet, P., Reed, S., Anguelov, D., . . . Ieee. (2015, Jun 07-12). Going Deeper with Convolutions.*IEEE Conference on Computer Vision and Pattern Recognition* [2015 ieee conference on computer vision and pattern recognition (cvpr)]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British journal of psychology*, *54*(2), 101-114.

Tanne-Gariepy, J., Rouiller, E. M., & Boussaoud, D. (2002). Parietal inputs to dorsal versus ventral premotor areas in the macaque monkey: evidence for largely segregated visuomotor pathways [Article]. *Experimental Brain Research*, *145*(1), 91-103. https://doi.org/10.1007/s00221-002-1078-9

Tarr, M. J., & Bulthoff, H. H. (1993). Conditions for viewpoint dependence and viewpoint invariance: What mechanisms are used to recognize an object? *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1496-1507.

Tarr, M. J., & Bulthoff, H. H. (1995). IS HUMAN OBJECT RECOGNITION BETTER DESCRIBED BY GEON STRUCTURAL DESCRIPTIONS OR BY MULTIPLE VIEWS - COMMENT ON BIEDERMAN AND GERHARDSTEIN (1993) [Article]. *Journal of Experimental Psychology-Human Perception and Performance*, *21*(6), 1494-1505. https://doi.org/10.1037/0096-1523.21.6.1494

Tarr, M. J., & Bulthoff, H. H. (1998). Image-based object recognition in man, monkey and machine [Article]. *Cognition*, *67*(1-2), 1-20. https://doi.org/10.1016/s0010-0277(98)00026-2

Tarr, M. J., & Bülthoff, H. H. (1999). *Object recognition in man, monkey, and machine* (Vol. 67). The MIT Press.

Tarr, M. J., & Hayward, W. G. (2017). The concurrent encoding of viewpoint-invariant and viewpoint-dependent information in visual object recognition [Article]. *Visual Cognition*, *25*(1-3), 100-121. https://doi.org/10.1080/13506285.2017.1324933

Tarr, M. J., & Pinker, S. (1989). MENTAL ROTATION AND ORIENTATION-DEPENDENCE IN SHAPE-RECOGNITION [Article]. *Cognitive Psychology*, *21*(2), 233-282. https://doi.org/10.1016/0010-0285(89)90009-1

Tarr, M. J., & Pinker, S. (1990). WHEN DOES HUMAN OBJECT RECOGNITION USE A VIEWER-CENTERED REFERENCE FRAME [Article]. *Psychological Science*, *1*(4), 253-256. https://doi.org/10.1111/j.1467-9280.1990.tb00209.x

Tjan, B. S. (2001). Adaptive object representation with hierarchically-distributed memory sites [Proceedings Paper]. *Advances in Neural Information Processing Systems 13*, *13*, 66-72.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015, Dec 11-18). Learning

Spatiotemporal Features with 3D Convolutional Networks.*IEEE International Conference on Computer Vision* [2015 ieee international conference on computer vision (iccv)]. IEEE International Conference on Computer Vision, Santiago, CHILE.

Ungerleider, L. G., & Mishkin, M. (1982). Two Cortical Visual Systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549-586). MIT Press.

Vannuscorps, G., Wurm, M. F., Striem-Amit, E., & Caramazza, A. (2019). Large-Scale Organization of the Hand Action Observation Network in Individuals Born Without Hands [Article]. *Cerebral Cortex*, *29*(8), 3434-3444. https://doi.org/10.1093/cercor/bhy212

Voci, A. (2003). *Processi psicosociali nei gruppi* (E. Laterza, Ed. 15th ed.). Gius. Laterza & Figli S.p.A.

Wang, X. L., Farhadi, A., & Gupta, A. (2016, Jun 27-30). Actions similar to Transformations.*IEEE Conference on Computer Vision and Pattern Recognition* [2016 ieee conference on computer vision and pattern recognition (cvpr)]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA.

Wohrer, A., & Kornprobst, P. (2009). Virtual Retina: A biological retina model and simulator, with contrast gain control [Article]. *Journal of Computational Neuroscience*, *26*(2), 219-249. https://doi.org/10.1007/s10827-008-0108-4

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., & Xiao, J. (2015). 3d shapenets: A deep representation for volumetric shapes. Proceedings of the IEEE conference on computer vision and pattern recognition,

Wurm, M. F., & Caramazza, A. (2019). Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language [Article]. *Nature*

Communications, *10*, 10, Article 289. https://doi.org/10.1038/s41467-018-08084-y

Wurm, M. F., & Caramazza, A. (2022). Two 'what' pathways for action and object recognition [Review]. *Trends in Cognitive Sciences*, *26*(2), 103-116. https://doi.org/10.1016/j.tics.2021.10.003

Wurm, M. F., Caramazza, A., & Lingnau, A. (2017). Action Categories in Lateral Occipitotemporal Cortex Are Organized Along Sociality and Transitivity [Article]. *Journal of Neuroscience*, *37*(3), 562-575. https://doi.org/10.1523/jneurosci.1717-16.2017

Wurm, M. F., & Lingnau, A. (2015). Decoding Actions at Different Levels of Abstraction [Article]. *Journal of Neuroscience*, *35*(20), 7727-7735. https://doi.org/10.1523/jneurosci.0188-15.2015

Xu, C., Xiong, C., & Corso, J. J. (2017). Action Understanding with Multiple Classes of Actors. *arXiv preprint arXiv:1704.08723*.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex [Article]. *Nature Neuroscience*, *19*(3), 356-365. https://doi.org/10.1038/nn.4244

Yang, Z. H., Zhao, X. Q., Wang, C. X., Chen, H. Y., & Zhang, Y. M. (2008). Neuroanatomic correlation of the post-stroke aphasias studied with imaging [Article; Proceedings Paper]. *Neurological Research*, *30*(4), 356-360. https://doi.org/10.1179/174313208x300332

Yin, R. K. (1969). Looking at upside-down faces. *Journal of experimental psychology*, *81*(1), 141.

Zhang, Z. L., & Sabuncu, M. R. (2018). Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *Advances in Neural Information Processing Systems 31 (Nips 2018)*, *31*.

Zhao, H., Torralba, A., Torresani, L., & Yan, Z. (2019). Hacs: Human action clips and segments dataset for recognition and temporal localization. Proceedings of the IEEE International Conference on Computer Vision,

Zhou, B. L., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452-1464. https://doi.org/10.1109/tpami.2017.2723009