Proceedings of the 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018

Selective Ensemble Learning based Human Action Recognition Using Fusing Visual Features

Chao Tang
Department of Computer Science and Technology
Hefei University
Hefei, China
tangchao77@sina.com

Rustam Stolkin
Extreme Robotics Lab (ERL),
School of Metallurgy and Materials
University of Birmingham
Birmingham, UK
r.stolkin@cs.bham.ac.uk

Chunling Hu
Department of Computer Science and Technology
Hefei University
Hefei, China
huchunling8787@aliyun.com

Huosheng Hu
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
hhu@essex.ac.uk

Xiaofeng Wang
Department of Computer Science and Technology
Hefei University
Hefei, China
xfwang@iim.ac.cn

Le Zou
Department of Computer Science and Technology
Hefei University
Hefei, China
zoule1983@163.com

***Abstract*—The selection of motion feature directly affects the recognition effect of human action recognition method. Single feature is often affected by human appearance, environment, camera settings and other factors, and its recognition effect is limited. This paper propose a novel action recognition method by using selective ensemble learning, which is a special paradigm of ensemble learning. Moreover, this paper presents a fast and efficient action description feature and a novel recognition algorithm. Robust discriminant mixed features are learnt from behavioral video frames as behavioral descriptors, The recogniton algorithm using selective ensemble learning can achieve fast classification. Experimental results show that the proposed method achieves ideal recognition results on the self-built indoor behavior data set and public data set.**

Keywords- human action recognition, visual features, selective ensemble learning, ensemble learning, features fusion

I. INTRODUCTION

Human action research has broad application prospects and great theoretical significance. From the perspective of application, as an important part of computer vision understanding and video analysis, human action recognition research results can be widely used in intelligent video surveillance[1, 2], human-computer interaction[3], video analysis[4] and video retrieval systems[5].Human action recognition research is devoted to extract high-level semantic knowledge from the underlying video or image data, so that the computer has the ability of human visual understanding. Action recognition consists of three key modules: (i) Action representation method corresponding to feature extraction module; (ii) Action modeling, namely, feature statistics model module; (iii) Action recognition/classification module.

Since the issue of human action recognition has been put forward, a large number of research methods and achievements have appeared in the academic circles [6-12]. In current research on human action recognition, it is typical that single feature is selected as the action representation, but single feature is also limited, e.g. modeling precision can be affected by the shadow and contour quality in shape features; background movement in action features can lead to bias; human body geometrical characteristics only adapt to the controlled environment and requires priori model knowledge. Thus, fusion features serving as action descriptors is an important direction of research. Maity et al. [13] developed a spatial-temporal body parts movement (STBPM)-based human action recognition method with which action recognition was realized using action-code classification (ACC). ACC was a code-based classification method that encoded each action by analyzing STBPM features without requiring training process. Chen et al. [14] put forward an action recognition deep learning framework with time invariance, which was found robust in recognizing the variation of movement velocity. The spatial and dynamic features were extracted from this key frame using convolutional neural networks (CNN), followed by fusion of both features in the convolutional fusion layer. A new kind of human action features that combined global action information and local action information was developed by Liu et al. [15]. Periodical actions served as the features in global action features, whereas enhanced histograms of action words (EHOM) were used in local action features. Human action recognition rate could be improved by the combination of both features.

In order to reduce the computational complexity of the algorithm, we use three simple computational but important features for human perceptual video behavior, namely RGB modal-based Fourier descriptors (RDB-FD), depth modal-based spatial-temporal interest point (D-STIPD), and joints modal-based joint point distribution feature (J-JPDF), as the description of video information. These features respectively reflect shape structure information, the motion information and depth information in video. These three features are very fast in computation and low in memory consumption. Finally, the three different features are combined as a mixed feature. The experimental results show that the robustness and recognition performance of the mixed feature are better than that of the single feature. A novel selective ensemble method is used for recognizing human action. Ensemble learning has attracted

much attention in machine learning circles because it can significantly improve the generalization ability of a learning system. However, with the increase of the number of basic learning machines, the prediction speed of ensemble learning machines decreases dramatically, and the required storage space is also rapid. The main purpose of selective ensemble learning is to further improve the predictive effect of ensemble learning machines, improve the predictive speed of ensemble learning machines, and reduce their storage requirements.

This paper is based on the channel feature information extracted by Kinect depth sensor for human antion recognition. Kinect is a somatosensory peripheral for the Xbox 360 that integrates a variety of perception technologies. The Kinect depth sensor consists of a RGB camera, a depth sensor and a multi-array microphone.

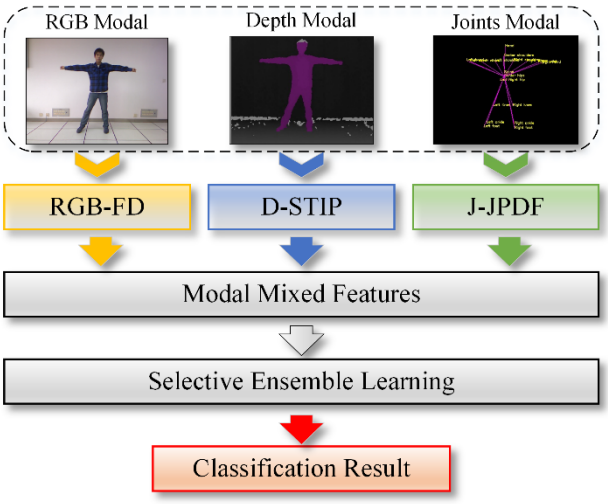

Figure 1. Human action recognition framework using selective ensemble learning method.

II. Overview of System

In order to enhance the system's robustness and practicability, and to exploit effectively the advantages of different features, this paper utilizes various modal data provided by the Kinect sensor and fuses three kinds of different features as the descriptors of the actions. The framework uses Fourier descriptor for RGB maps, spatial temporal interest point for depth maps, and joint point distribution feature for data from joints. After that, a selective ensemble learning algorithm is adopted for the classification. The system flowchart is illustrated in Fig1. The method maintains the computational convenience from simple features while ensuring the robustness and recognition capability of the features at the same time.

III. Feature Expression

The purpose of action feature extraction is to extract the feature information from the underlying data to characterize human motion.The effect of action feature extraction has an important influence on human action recognition. In Section III, three different modal representation methods are introduced.

A. RGB modal-based Fourier Descriptors

Fourier descriptor based on object coordinate sequence has the best shape recognition performance. Four shape expressions can be derived from the coordinates of the closed boundary points in the image plane, which are curvature function, centroid distance, complex coordinate function and chord length function. In order to avoid complex computation, Fourier descriptors based on center distance are used to process RGB modal image data. The illustration of the centroid distance of human body in the figure 2. Firstly, the region of interest of human body is detected, then the human boundary curve is obtained by boundary tracking algorithm. In the spatial domain, the centroid distance $R(l)$ is defined as the distance from the body boundary point (x_l, y_l) to the center point of the human body (x_c, y_c) :

$$R(l) = \sqrt{(x_l - x_c)^2 + (y_l - y_c)^2} \tag{1}$$

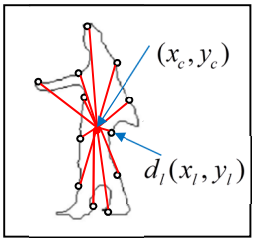

Figure 2. Illustration of the centroid distance of human body.

After Fourier transform, the centroid distance function in spatial domain can be described by Fourier transform coefficients in frequency domain. For the centroid distance function, only the positive frequency axis is considered, because the Fourier transform coefficients F_i are symmetrical:

$$| F_{-i} | = | F_i | \tag{2}$$

Therefore, the shape descriptors derived from the centroid distance can be obtained as follows:

$$FD = \left[\frac{|f_2|}{|f_1|}, \frac{|f_3|}{|f_1|}, \ldots, \frac{|f_{N-1}|}{|f_1|} \right] \tag{3}$$

In order to ensure that all the shape features of objects have the same length, the number of all boundary points must be unified before Fourier transform is implemented, so that fast Fourier transform can be used to improve the efficiency of the algorithm. Therefore, we can get the feature of RGB-FD, as shown in the formula.

$$RGB - FD = [FD_1, FD_2, \ldots, FD_M] \tag{4}$$

B. Depth modal-based STIP

In recent years, researchers have proposed an action representation and recognition method based on spatio-temporal interest points. This method first detects interest points which represent spatio-temporal events in video sequences, then extracts various features of spatio-temporal interest points to represent actions, and finally establishes an action classifier for action recognition. Laptev et al. extend the

2D Harris corner[16] to the 3D Harris corner point[17], which is a significant change point in the spatio-temporal domain. Laptev et al. performed an admirable research on spatiotemporal domain for action recognition. Still, 3D counterpart of 2D methods is inadequate for action recognition in spatio-temporal domain. Dollar [18] proposes a Gabor filter based spatio-temporal interest point detection algorithm, and proposes a feature description operator based on cuboids. PCA algorithm is used to reduce the dimension of features. Finally, the nearest neighbor classifier based on χ^2 distance is used for action recognition.

For gray-scale images, the feature of spatio-temporal interest points is that the brightness of pixels in the image changes dramatically in both spatial and temporal dimensions simultaneously. The spatio-temporal interest points of depth information are pixels whose depth values change dramatically in both time and space domains.Consequently, we propose a novel local depth-based STIP to represent human action.Therefore we can get the feature of depth-STIP, as shown in the formula.

$$D-STIP=[STIP_1,STIP_2,...,STIP_M] \quad (5)$$

C. Joints modal-based JPDF

In recent years, Kinect camera has been widely used for its high resolution and low cost. Kinect can not only collect color images and depth images, but also extract human skeleton information.3D skeleton information is the coordinates of the three-dimensional position of the joints in the human body. It can also determine the various parts of the human body, such as the hand, head and body, and determine their specific three-dimensional position information.When a person walks into Kinect's field of view, Kinect can use tracking technology to find the position of the user's 20 joint points, which are represented by three-dimensional coordinates (x,y,z) The skeleton joint diagram obtained by Kinect is shown in Figure 3.

Human action recognition based on 3D skeleton is superior to action recognition based on depth data in many aspects. For example, in terms of computing speed, human action recognition based on 3D skeleton is obviously faster and more efficient, because the amount of skeleton data is much smaller than the amount of image data, and it can extract a more concise feature description.

In this paper, we propose a skeletal feature based on geometric characteristics, which is called the joint position distribution feature. This feature describes the motion state of human body by the spatial position between the joint points and other information. The advantage of this feature is that it not only pays attention to the details of human motion, but also captures the spatial information of bones.

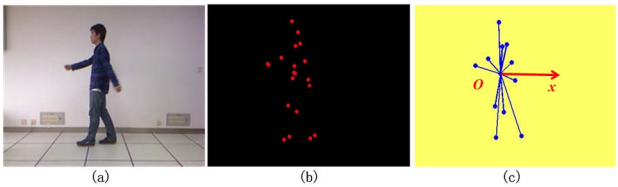

Figure 3. (a) Original image (b) Joints of the human body (c) Polar coordinate of joints of the human body skeleton

In the plane take the human body's central joint point O, called the pole, a ray Ox, called the polar axis, and then select a length unit and angle of the positive direction (usually counterclockwise direction). For any joint point M in the plane, the length of the line segment OM is represented by r, and the angle from Ox to OM is represented by θ. r is called the polar radius of point M, θ is called the polar angle of point M, and the pair of ordinal numbers (r, θ) is called the polar coordinates of point M. Thus the coordinate system is called the polar coordinate system.

Therefore, we can get the relative position feature of human 22 joint points based on joint modal data as follows:

$$J-JPDF=[J-JPDF_1,J-JPDF_2,...,J-JPDF_{22}] \quad (6)$$

D. Features Fusion

RGB-FD, D-STIP and J-JPDF are extracted from single frame images. In order to improve the accuracy of action recognition, RGB-FD, D-STIP and J-JPDF are combined to form a mixed eigenvector RDJ_t, as shown in Eq. (9).

$$RDJ_t=[RGB-FD_t,D-STIP_t,J-JPDF_t] \quad (7)$$

IV. METHOD

Compared with traditional ensemble learning, the selective ensemble learning can generate the base classifiers with stronger generalization ability. In this paper, we propose a selective ensemble classification method using multi-modal image information to recognize human action.We choose the best integrated classifier and obtain the final result by majority vote to recognize human action. The flow chart of the selective ensemble method has been shown in Figure 4.

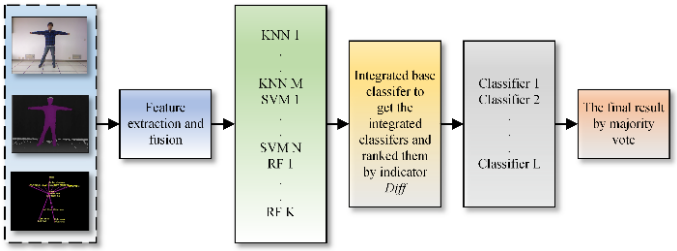

Figure 4. Te flow chart of the selective ensemble method

We choose the KNN, SVM, and RF as the base classifier because they are implemented simply and recognize action effectively. The key in ensemble learning is the accuracy and diversity of base classifiers.

To measure ensemble diversity, a classical approach is to measure the pairwise similarity/dissimilarity between two classifiers, and then average all the pairwise measurements for the overall diversity. Given a data set $Data=\{(x_1,y_1),...,(x_m,y_m)\}$, for binary classification (i.e., $y_i\in\{-1,+1\}$), we have the following contingency table for two classifiers h_i and h_j, where $a+b+c+d=m$ are non-negative

variables showing the numbers of examples satisfying the conditions specified by the corresponding rows and columns.

TABLE I. CLASSIFIERS RELATIONAL TABLE OF CLASSIFICATION OF THE SAMPLES

	$h_i = +1$	$h_i = -1$
$h_j = +1$	a	c
$h_j = -1$	b	d

In this paper, Disagreement Measure (DM) will be introduced based on these variables.

$$\begin{cases} DM_{i,j} = \dfrac{b+c}{a+b+c+d} \\ DM_i = \dfrac{\sum_{2}^{L} DM_{i,j}}{L-1} \end{cases} \tag{8}$$

We use the indicator *Diff* to rank the single classifier which is based on the accuracy and its diversity. *Diff* can be described as follows:

$$Diff_i = \alpha_1 * Acc_i + \alpha_2 * DM_i \tag{9}$$

Acc_i is the accuracy of the single classifier; DM_i can reflect the diversity of the classifier. The larger $Diff_i$ is, the better the diversity the classifier will be. The different α can impact the choice of the classifier ($\alpha_1 = 0.4$, $\alpha_2 = 0.6$).

V. EXPERIMENTAL RESULTS

In order to verify the validity of the proposed algorithm, a large number of comparative experiments have been done on the self-built indoor behavior data set, the open Gaming 3D video database and the CAD-60 database.

A. Human Action Datasets

We built an indoor action dataset through the Kinect sensor. Microsoft Kinect's RGB-D sensor is a peripheral device designed for Microsoft X-BOX video games. Figure 5 shows some example visual clips of the 10 types of human actions.

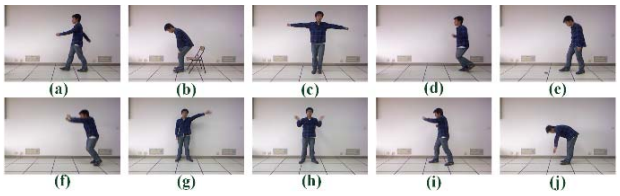

Figure 5. Ten sample actions from Kinect dataset

Gaming 3D dataset (G3D) captured by Kingston University in 2012 focuses on real-time action recognition in gaming scenario. It contains 10 subjects performing 20 gaming actions. Figure 6 shows some example visual clips of the 20 types of human actions.

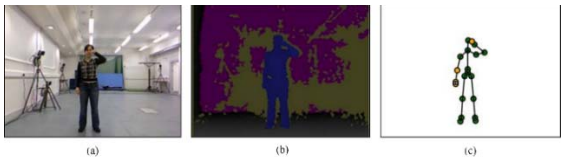

Figure 6. Example images for G3D dataset. (a) RGB image (b) depth image (c) Joints imag

CAD-60 dataset was captured by Cornell University in 2011, motivated by the fact that true daily activities rarely occur in structured environments. Figure 7 shows some example visual clips of the 12 types of human actions.

Figure 7. Example images for CAD-60 dataset.

B. Performance Evaluation

We first test the effectiveness of representing actions from the features of RGB-FD+D-STIP+J-JPDF. This features mixed three kinds of descriptors: RGB-FD, D-STIP and J-JPDF. In the experiment, we randomly divided all the videos into training data and test data at a ratio of 2:8 each time. The final test result is the average of the 10 test results.Figure 8-10 show the accuracy of the proposed method in single-frame action recognition in the form of confusion matrix. Among them, the (i, j) element of the confusion matrix represents the proportion of class i action classified as class j action, that is, the greater the value of the diagonal, the better the classification effect.

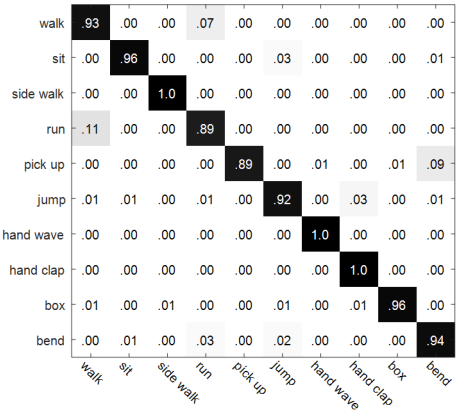

Figure 8. Confusion matrix of the proposed action recognition using Kinect dataset.

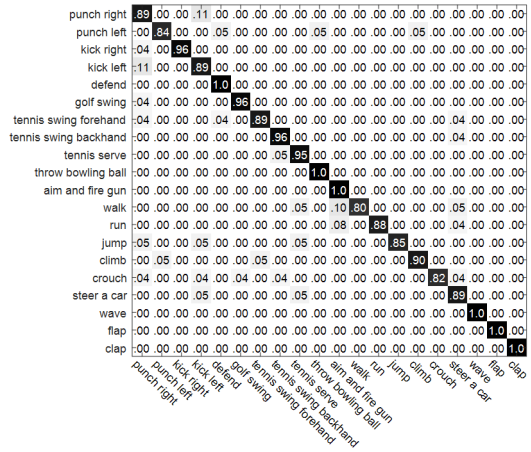

Figure 9. Confusion matrix of the proposed action recognition using G3D dataset.

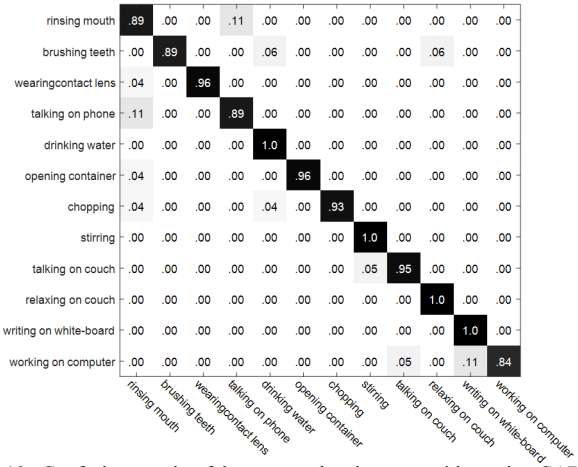

Figure 10. Confusion matrix of the proposed action recognition using CAD-60 dataset.

Figure 8 shows that the average accuracy is about 95% for selective ensemble learning strategy using multi-modal features. From Figure 8, we can see that recognition result for action ‘run’ is not good. We analyze the reasons that in Figure 8 action ‘run’ is mainly classified to action ‘walk’. It can be explained that the action 'run' and the action 'walk' are very similar to that in a single frame. Meanwhile, because the action ‘pick up’ and ‘bend’ are also similar, recognition rate of ‘pick up’ and ‘bend’ are not high. To solve this problem, we need to extract more effective features to distinguish these types of action. We noticed that the three actions of side walk, hand wave and hand clap achieved 100% recognition rate.Figure 9 and Figure 10 show the experimental results under the G3D and CAD-60 datasets, from which we can find that the proposed method has achieved good recognition rate.

C. *Comparative Results*

In Table 2, we compare the recognition rates using mixed features and single feature in three different datasets.From the comparison of the experimental results, we can see that the recognition accuracy using combined features, RGB-FD+D-STIP+J-JPDF, is obviously better than that using single feature. The effectiveness of the proposed feature is fully verified.Therefore, the fusion of different modes of human action feature is more conducive to characterize human and, with high discrimination and distinction.

TABLE II. TABLE I. THE RECOGNITION RATE OF HUMAN ACTION USING DIFFERENT FEATURES

Datasets	**Descriptors**	***Precision***	***Recall***	***F-Measure***
Kinect Dataset	RGB-FD+D-STIP+J-JPDF	0.95	0.95	0.95
	RGB-FD+D-STIP	0.91	0.91	0.91
	RGB-FD+J-JPDF	0.92	0.92	0.91
	D-STIP+J-JPDF	0.90	0.90	0.906
	RGB-FD	0.86	0.86	0.86
	D-STIP	0.70	0.70	0.69
	J-JPDF	0.89	0.89	0.89
G3D Dataset	RGB-FD+D-STIP+J-JPDF	0.92	0.92	0.92
	RGB-FD+D-STIP	0.87	0.86	0.87
	RGB-FD+J-JPDF	0.85	0.85	0.85
	D-STIP+J-JPDF	0.87	0.86	0.87
	RGB-FD	0.80	0.80	0.80
	D-STIP	0.70	0.70	0.69
	J-JPDF	0.85	0.85	0.85
CAD-60 Dataset	RGB-FD+D-STIP+J-JPDF	0.94	0.94	0.94
	RGB-FD+D-STIP	0.88	0.87	0.88
	RGB-FD+J-JPDF	0.85	0.85	0.85
	D-STIP+J-JPDF	0.85	0.86	0.86
	RGB-FD	0.80	0.81	0.81
	D-STIP	0.70	0.70	0.69
	J-JPDF	0.86	0.86	0.86

VI. Conclusion

In this paper, we propose a new selective ensemble method to recognize human action by combining RGB modal information, depth modal information and joints modal information. When generating the integrated classifier, we choose the suitable base classifier using new indicator *Diff* . Then we test our method in three kinds of multi-modal datasets containing our Kinect dataset, G3D dataset and CAD-60 dataset. The selective ensemble method is efficient in recognition human action; and we can obtain an accuracy of 95%. In the future work, we will balance the complexity and accuracy, and give a better adaptive selective ensemble framework.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (61672204), the Scientific Research Fund Project of Talents Hefei University (15RC07), the Excellent Talents Training Funded Project of Universities of Anhui Province (gxfx2017099), the Key Scientific Research Foundation of Education Department of Anhui Province (KJ2014A212, KJ2018A0555) , the Visiting Scholar at Home and Aboard Funded Project of Universities of Anhui Province (gxfxZD2016209) and the Grant of Major Science and Technology Project of Anhui Province(17030901026)

References

[1] A. Ladjailia, I. Bouchrika, N. Harrati, and Z. Mahfouf, *Encoding Human Motion for Automated Activity Recognition in Surveillance Applications* (Computer Vision: Concepts, Methodologies, Tools, and Applications). Hershey, PA, USA IGI Global, 2018, pp. 2042-2064.

[2] K.-E. Ko and K.-B. Sim, "Deep convolutional framework for abnormal behavior detection in a smart surveillance system," *Engineering Applications of Artificial Intelligence,* vol. 67, pp. 226-234, 2018.

[3] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review,* vol. 43, no. 1, pp. 1-54, 2015.

[4] S. Nazir, M. H. Yousaf, J.-C. Nebel, and S. A. Velastin, "A Bag of Expression Framework for Improved Human Action Recognition," *Pattern Recognition Letters,* vol. 103, pp. 39-45, 2018.

[5] X. Liu, G.-F. He, S.-J. Peng, Y.-m. Cheung, and Y. Y. Tang, "Efficient Human Motion Retrieval via Temporal Adjacent Bag of Words and Discriminative Neighborhood Preserving Dictionary Learning," *Ieee Transactions on Human-Machine Systems,* vol. 47, no. 6, pp. 763-776, Dec 2017.

[6] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications,* vol. 76, no. 3, pp. 4405-4425, 2017.

[7] L. Lo Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," (in English), *Pattern Recognition,* Article vol. 53, pp. 130-147, May 2016.

[8] D. Das Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector," *Visual Computer,* vol. 32, no. 3, pp. 289-306, 2016.

[9] M. Ziaeefard and R. Bergevin, "Semantic human activity recognition: A literature review," *Pattern Recognition,* vol. 48, no. 8, pp. 2329-2345, Aug 2015.

[10] Z. Zhang, S. Liu, S. Q. Liu, L. Han, and Y. X. Shao, "Semantic Analysis in Human Action Recognition: A Comprehensive Study," in *Proceedings of the Third International Conference on Communications, Signal Processing, and Systems*, 2015, vol. 322, pp. 573-580, Switzerland: Springer, Cham.

[11] C. H. Lim, E. Vats, and C. S. Chan, "Fuzzy human motion analysis: A review," *Pattern Recognition,* vol. 48, no. 5, pp. 1773-1796, May 2015.

[12] C. L. He and W. Pan, "Study on human Action Recognition Algorithms in videos," in *Proceedings of the 2015 International Symposium on Computers & Informatics*, 2015, vol. 13, pp. 703-709.

[13] S. Maity, D. Bhattacharjee, and A. Chakrabarti, "A Novel Approach for Human Action Recognition from Silhouette Images," *IETE Journal of Research,* vol. 63, no. 2, pp. 160-171, 2017.

[14] H. Chen, J. Chen, R. Hu, C. Chen, and Z. Wang, "Action recognition with temporal scale-invariant deep learning framework," *China Communications,* vol. 14, no. 2, pp. 163-172, 2017.

[15] H. Liu, Z. Ju, X. Ji, C. S. Chan, and M. Khoury, "Study of human action recognition based on improved spatio-temporal features," in *Human Motion Sensing and Recognition*: Springer, 2017, pp. 233-250.

[16] C. Harris, "A combined corner and edge detector," in *Proc of the 4th Alvey Vision conference*, 1988, vol. 15, no. 50, pp. 147-151, 1988.

[17] I. Laptev and T. Lindeberg, "Space-time interest points," in *the 9th IEEE International Conference on Computer Vision*, Nice, France, 2003, pp. 432-439, New York, NY, USA: IEEE, 2003.

[18] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Beijing, China, 2005, pp. 65-72, New York, NY, USA: IEEE, 2005.