IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.DOI

A Survey of Deep Learning Solutions for Multimedia Visual Content Analysis

MUHAMMAD SHAHROZ NADEEM[1], VIRGINIA N. L. FRANQUEIRA[1] (Member, IEEE), XIAOJUN ZHAI[2] (Member, IEEE), and FATIH KURUGOLLU[1] (Senior Member, IEEE)

[1]College of Engineering and Technology, University of Derby, Derby, DE22 3AW, United Kingdom
[2]School of Computer Science and Electronics Engineering, University of Essex, Colchester, CO4 3SQ, United Kingdom

Corresponding author: Muhammad Shahroz Nadeem, (m.nadeem@derby.ac.uk)

ABSTRACT The increasing use of social media networks on handheld devices, especially smartphones with powerful built-in cameras, and the widespread availability of fast and high bandwidth broadband connections, added to the popularity of cloud storage, is enabling the generation and distribution of massive volumes of digital media, including images and videos. Such media is full of visual information and holds immense value in today's world. The volume of data involved calls for automated visual content analysis systems able to meet the demands of practice in terms of efficiency and effectiveness. Deep Learning (DL) has recently emerged as a prominent technique for visual content analysis. It is data-driven in nature and provides automatic end-to-end learning solutions without the need to rely explicitly on predefined handcrafted feature extractors. Another appealing characteristic of DL solutions is the performance they can achieve, once the network is trained, under practical constraints. This paper identifies eight problem domains which require analysis of visual artefacts in multimedia. It surveys the recent, authoritative, and best performing DL solutions and lists the datasets used in the development of these deep methods for the identified types of visual analysis problems. The paper also discusses the challenges that DL solutions face which can compromise their reliability, robustness, and accuracy for visual content analysis.

INDEX TERMS Visual Content Analysis, Deep Learning, Machine Learning, Dataset.

I. INTRODUCTION

IN recent years, the availability of handheld devices with high storage capacity (complemented by the cloud) and with integrated cameras has caused a boom in the generation of digital media (images and videos) by individuals. Such content is vastly shared through high bandwidth and fast broadband connections, helped by the reaching power of social media. It has been estimated that there were about 4 trillion images worldwide stored on devices, storage media, and in the cloud by 2016 and that, in 2020 alone, 1.4 trillion new digital photographs will be captured worldwide [1]. Following a similar trend, it has been reported (in July 2015) that more than 400 hours of video were uploaded to YouTube every minute [2]. Adding to the phenomena is the increasing deployment of CCTV cameras, capturing high volumes of media in public and private spaces, to enhance security and prevent crimes [3].

Deep Learning (DL) has been proven to be effective at processing and analysing visual media. It has the ability to extract and learn abstract information compared to shallow methods [4]. DL methods eliminate the need for handcrafted feature extraction and representation [5]. This enables it to take advantage of increasing computational power and data without the involvement of domain experts [6]. In DL, feature extraction and classification are combined together during training in an end-to-end manner [7]. Usually, training a deep network is not easy and is time demanding. However, once trained, deep methods can then process data in seconds. These advantages of DL make it an attractive option for visual content analysis.

There are many high-quality in-depth surveys for specific problems in visual content analysis (e.g., [8]–[10]). They present deep architectures and solutions focusing on a particular visual task. However, no survey provides an overview of DL applied across different problem domains related to visual content analysis, although it often happens that solutions from one such problem domain can be re-applied or adapted to another (e.g., [11]–[14]). Therefore, this survey aims to fill this gap, and develop an understanding of the critical aspects of DL methods that enhance content analysis through visual

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

artefacts.

In summary, the main contributions of this paper are the following.

1) Survey of DL based solutions for eight classes of visual content analysis problems.
2) Compilation of datasets that have been used to develop deep methods for each identified class of problems.
3) Review of limitations of DL solutions that could have a negative impact on the deep methods.

The remaining of this paper is organised as follows. Section II provides an overview of the background on the most prominent types of DL methods. Section III identifies eight classes of problems related to visual content analysis, and surveys DL-based solutions for them. Section IV presents a compiled list of authoritative and recent datasets relevant to the surveyed solutions. Section V focuses on the shortcomings of DL methods and elaborates on future research directions. Finally, Section VI concludes the paper.

II. BACKGROUND

DL is the most attractive branch of Machine Learning (ML) techniques, which is being actively utilised to extract high-level features to model abstract concepts. Most DL methods are based on the supervised learning strategy. However, the hectic process of labelling and developing large scale dataset is costly and requires ample amount of manpower and effort. DL methods are moving towards other forms of learning, which include semi and unsupervised approaches. Reinforcement learning is another interesting strategy to train DL methods through interaction with the environment.

Traditionally, ML approaches used carefully designed feature extractors which required domain knowledge [4], [6]. These handcrafted features, limited in capacity, often failed in unforeseen real-life scenarios. DL, inspired by the human nervous system, is a subset of ML. It is data-driven and able to learn abstract and complex features automatically. However, training deep networks is hard and the two prerequisites for training are high computational power and a huge volume of data. The re-emergence of DL surfaced when AlexNet [15] won the 2012 ImageNet competition. This network contained 5 convolutional and 3 fully connected layers. The convolution layers were followed by ReLU non-linearity and max pooling. Containing a total of 650,000 neurons and 60 million parameters took 2 days to train. In the majority of the DL methods convolution layer is the main workhorse used as a discriminator and feature extractor. This layer is responsible for learning low or high-level features. Due to this reason, Convolutional Neural Networks (CNN) are the most popular and commonly used DL networks.

After AlexNet [15], many new CNN architectures were designed for image classification. These included ZFNet [16], VGGNet [17], GoogLeNet [18] and ResNet [19], which have overcome human performance for the same benchmark. Rather than just stacking layers to make the networks deep, different designs were introduced. ZFNet [16] used deconvolution techniques to visualise the learned features at different levels of depth. Inception module [18] was designed to make the network wider, deeper and computationally less expensive. However, their network suffered from the vanishing gradient problem. Gradients are very essential for learning and are at the core of backpropagation. Where gradients are passed backwards during training to update the learned features, however in deeper networks the gradients became so small that they eventually became zero. This problem was solved in ResNet [19], by the addition of skip connections. This enabled training far deeper networks possible.

Deep networks are often comprised of individual components and layers, each serves a different purpose. Deep networks not only suffer from vanishing gradients, but they are also prone to over-fitting. This kills the ability of the network to generalise. Different approaches are used to avoid over-fitting, they include: applying regularisation on the loss function, adding weight decay, dropout layer [20], normalisation (batch or instance) layers, or simply early stopping strategy. These factors make designing and training a deep network difficult because of too many hyper-parameters and design choices that vary for distinct problems. The most prominent DL methods are based on CNN, other then them Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM) networks, Auto-Encoders and Generative Adversarial Networks (GAN) are also heavily used [21].

Figure 1 shows the most popular types of DNN, we have categorised these networks as Feedforward, Recurrent and Generative Adversarial. MLP is the traditional feedforward networks, however, now they are heavily overshadowed by CNN. These networks do not have cyclic connections between them, information is passed forward and gradients are passed backwards. The most recent addition to the feedforward network is the Capsule Networks (CapsNets), designed specifically to remove the inherent limitations of CNN's as a discriminator. Capsules are trained using a dynamic routing algorithm and are the focus of the current research in DL. Still, in its infancy, CapsNets are an active area of research in DL, are shown to have the potential to change DL landscape. Finally, auto-encoders consist of an encoder and a decoder network. The encoder converts the input data to an intermediary representation also known as latent variables. The decoder reconstructs the input samples from these latent variables. A unique characteristic of auto-encoders is the presence of a bottleneck. The simplest way to create a bottleneck is by restricting the number of hidden neurons. The input is then passed through this bottleneck and a compressed structural representation of data is learned. However, one must be careful that the network does not memorise the data. Yet should learn features that accurately describe it. Types of auto-encoders include sparse, contractive, denoising and variational.

Feedforward network treat data samples independently thus no cyclic connections are present. In contrast to this, recurrent networks have cyclic connections as the data samples are not independent. The recurrent network takes into con-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

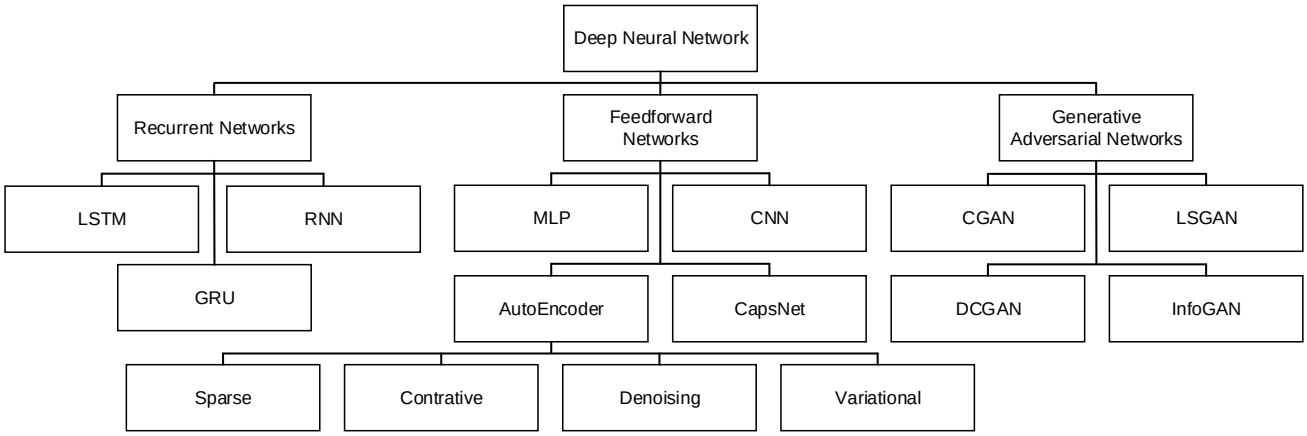

FIGURE 1. The most prominent types of DL techniques

sideration the learned experience during the training process. Due to which, they are used to handle temporal and sequential data e.g. Natural Language Processing (NLP) and Video Analytics. Recurrent networks consist of RNN, LSTM, and Gated Recurrent Unit (GRU). RNN is the most basic form of a recurrent network. In addition to the current input, the previous output is also passed back to neurons. This acts as memory however, RNN suffers from the vanishing gradient problem and cannot retains long term memory. LSTM network was designed to solve this problem. A typical LSTM cell takes three distinct inputs, the hidden state (also called short memory), cell state (also called long memory) and the current input at the time. These inputs are passed through three gates in the LSTM, the input, output and forget gate. The three inputs are updated individually and passed forward to the next LSTM cell. Gradients are passed through these gates which keeps the gradients from dying. Similar, gating technique is used by GRU to stop gradients from vanishing. However, in contrast to LSTM, GRU only has two gates called reset and update gates.

Relatively recent, GANs have been developed as a combination of a discriminator and generator networks. The two networks are put together in a competitive environment against each other. Through this adversarial relationship, both networks improve each other's performance. The generator network produces e.g. fake images, while the discriminator network tries to distinguish between the real and fake images. Over several iterations a generator network, starts to produce very realistic images which the discriminator find hard to distinguish, is obtained. GAN have the capability to produce photo realistic images from random noise, which makes them very interesting and a promising DNN. The most prominent types of GANs are the Conditional GAN (CGAN), Deep Convolution GAN (DCGAN), Least Square GAN (LSGAN) and InfoGAN. Amongst these categories, the most popular one is the DCGAN.

However, it must be emphasised that training and designing these deep networks is not an easy task as they involve hyper-parameter tuning, selection of the right evaluation criterion, activation and loss functions.

III. DL SOLUTIONS FOR VISUAL CONTENT ANALYSIS PROBLEMS

This section identifies eight classes of problems which require visual content analysis. Sections III-A to III-H survey DL solutions to address each of those classes.

A. ACTION RECOGNITION

Humans have a natural ability to recognise and interpret actions they are exposed to on a daily basis or rarely. However, to develop the same capability for machines is challenging. As most actions span over a certain time-frame, an understanding of temporal and motion components is required. Therefore, action recognition methods tilt towards videos datasets. A very recent survey on action recognition by Herath et al. [8] provides more specific information on challenges, proposed methods, and related datasets. Potential areas of application include smart video surveillance, video indexing and retrieval, autonomous driving, the gaming industry, and smart rehabilitation [22]. The most notable methods that have been proposed for action recognition are listed in Table 1.

One of the first attempts at using DL-based ConvNet for action recognition was by Simonyan et al. [23]. They proposed a Two-Stream CNN architecture; one stream processes the spatial component while the other stream processes the temporal component of videos. Feichtenhofer et al. [30] proposed a new model where they combined the spatial and temporal streams at different fusion levels, while improving the state of the art performance. They presented two novel convolutional and temporal fusion layers which are used to fuse the temporal and spatial components. Sun et al. [25] proposed another CNN architecture that factorises the learning of spatiotemporal components. Their method sequentially learnes 2D spatial kernels and 1D temporal kernel which result in significant gains in computational cost. Yue-Hei Ng et al. [26] investigate the incorporation of action information over a longer period of time by feeding the CNN learned

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

TABLE 1. DL solutions for Action Recognition

Method	Model	Media	Year	Dataset Used
Two-Stream ConvNet [23]	CNN	Video	2014	UCF-101 & HMDB51
Spatiotemporal ConvNet [24]	CNN	Video	2014	UCF-101 & Sport-1M[1]
Factorized ConvNet [25]	CNN	Video	2015	UCF-101 & HMDB51
Yue et al. [26]	LSTM-CNN	Video	2015	UCF-101 & Sport-1M[1]
C3D [27]	3D-CNN	Video	2015	UCF-101, Sport-1M[1] & ASLAN
Composite LSTM Model [28]	LSTM	Video	2015	UCF-101, HMBD-51 & Sport-1M[1]
LRCN [29]	LSTM	Image	2015	UCF-101
Two-Stream Fusion ConvNet [30]	CNN	Video	2016	UCF-101 & HMDB51
Dynamic Image Networks [31]	CNN	Image	2016	UCF-101 & HMDB51
ST-ResNet [13]	CNN	Video	2016	UCF-101 & HMDB51
LTC-CNN [32]	CNN	Video	2018	UCF-101 & HMDB51

features to an LSTM for ordered sequence modelling. Tran et al. [27] proposed a 3D CNN, named C3D. The authors claimed that C3D is better at video analysis tasks then 2D CNN's. A slightly different approach has been proposed by Bilen et al. [31]; they create dynamic images for video analysis from video samples by applying a "weighted average over time" approach. Such dynamic images capture the temporal and motion information inside the image. Varol et al. [32] have proposed a Long-term Temporal Convolutions (LTC) that takes advantage of long temporal structure.

Even though CNNâĂŹs are very good at learning generic representations, the same level of performance has not been observed for action recognition tasks. The main reason is been that CNN cant process temporal information which needs to be considered while learning video representation features. This information, in many action categories, spans over many seconds of a video sequence, therefore, temporal information has to be considered and preserved over a longer context. As a consequence, many proposed methods see a boost in their recognition performance after they are combined with handcrafted features [13]. The literature also shows that, in many cases, combining the proposed methods with Improved Dense Trajectory (IDT) [33] boosts the performance. However, in our view, action recognition is a well researched area.

B. VIOLENCE DETECTION

Violence is a subjective matter, thus not easy to define [34]. Primarily, violence detection methods make use of visual information. In many cases adding audio information also helps to improve detection performance [35], as certain sounds can be attributed to certain acts of violence, e.g., gun shots, screams, knocking, and yelling. Violence detection can be used in different public places like pubs, prisons, train and bus stations as well as public events such as concerts, sporting events, and protests [36].

Typically, violence has been considered as a sub-category of action related tasks [37], [38], and remained relatively underexplored [39]. Due to this reason, most of the action-based methods have been applied to violence tasks, with special interest towards violence detection among individuals and crowds [12]. However, there are certain aspects of violence tasks which make them different from action tasks [39].

Recently, the MediaEval Violent Scene Detection challenge (VSD 2014,2015), revived the interest in violence detection. The participating teams proposed many methods comprising traditional ML with handcrafted features, DL, and hybrid combining both approaches to yield the highest performance in order to win the challenge. Dai et al. [40] proposed a Deep Neural Network based method that fuses together multiple features to perform classification. They extracted audio-visual features of three types which include IDT, Space-Time Interest Points (STIP) and Mel-Frequency Cepstral Coefficient (MFCC). They later introduced another DL based method [41], in where they use a Two-Stream CNN, consisting of a spatial and temporal CNN. Followed by an LSTM on top of them to incorporate the features. In addition to these DL features, traditional audio-visual features were also added for violence detection. According to Dai et al. [41], DL features benefit by combining them with traditional features. Vu Lam et al. [42] and Marin et al. [43] came to the same conclusion that combining traditional and deep learned features improves performance.

Recently violence detection in a crowded scene has gained interest. Marsden et al. [12] proposed a residual DL-based crowd analysis system called ResnetCrowd. They use a dataset which was annotated for crowd counting, density level estimation, and violent behaviour recognition. A violence classification strategy based on DL solution was proposed by Peixoto et al. [35]. Their solution consists of a multi-task CNN network where each branch of the network is dedicated to a single violence category, therefore, each branch only learns features for that specific category. All the branches are combined and passed through an SVM to predict the final class. They use the Temporal Robust Features (TRoF) detector to produce three types of combination for motion images which are fed to the networks in addition to the still images. A list of methods for violence detection is shown in Table 2.

Violence detection is slightly less researched as compared to action recognition. However, in our view, certain aspects in

[1] https://cs.stanford.edu/people/karpathy/deepvideo/classes.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

a violent scenario make it distinct to normal action categories. Often a typical violent scene requires the participation of at least two individuals. The presence of blood, wounds, weapons, screams or shouts and possible loss of life or permanent disability requires that violence detection should be given special attention. There is a lack of comprehensive datasets for violence detection, due to ethical and moral considerations such data is also not publicly available.

C. PORNOGRAPHY DETECTION

The degree of acceptable sensuality differs amongst communities and cultures. However, pornography is unanimously regarded as unethical and immoral. Pornographic content can be captured by images, videos, animations, drawings and, more recently, by virtual reality [45]. Adult content filtering has huge application potential [46].

The first step in detecting pornographic content is the detection of nudity [47]. The major challenge in nudity detection is its subjective nature. There are different activities in which individuals show a lot of skin and perform certain actions that are not necessarily pornographic, e.g. swimming, wrestling, and sunbathing. However, pornography is a step further from nudity, where a single person or multiple individuals indulge in sexual activity, and this adds more complexity to automatic detection. Pornography involves the presence of sexual paraphernalia, and manifests in different categories and forms.

Due to these challenges, DL is a promising solution direction, and many methods have been proposed. Table 3 lists the most notable ones for pornography detection.

Moustafa [5] experimented with the existing DL based architectures that were making headlines in object detection in 2015. Specifically, he uses AlexNet and GoogLeNet and created a combined method, named AGNet. The study shows that AGNet had the best performance on the NPDI pornographic dataset, compared to existing methods at the time. They used keyframe images as input to the proposed network, however, the difference between the performance of the combined method and of the separate networks was not much significant.

Perez et al. [47] proposed a Two-Stream deep network, where they add motion information to a CNN. The techniques use for capturing motion information were optical flow and MPEG motion vectors. In their approach, they also evaluate the effects of early, mid-level and late fusion of static and dynamic information. Pornography-800 and Pornography-2K dataset are used to train the deep networks. They compared their model with third-party tools, traditional Bag-Of-Word (BOW) methods, and spatiotemporal networks, and show that DL outperformed the others. Late fusion of features performs consistently well on both datasets. However, they also made note that just using static images in CNN produced competitive results but the addition of motion information helps boost the performance of the DL method.

ACORDE is another method proposed by Wehrmann et al. [50] that combined CNN and LSTM to classify pornographic content into hard non-adult, easy non-adult or adult categories. ACORDE's CNN part extracts features, whereas its LSTM part focuses on sequence learning; the authors concluded that LSTM helped in video classification.

Varges da Silva et al. [51] experimented with spatiotemporal CNN networks using VGG-C3D CNN and ResNet R(2+1)D CNN. They compared performance to other CNN based methods for pornography detection, without combining them with other motion features such as optic flow or IDT.

Specialised image-based methods for pornographic detection using DL include the following. Wang et al. [48] proposed a Strongly-supervised Deep Multiple Instance Learning (SD-MIL) which they claimed to be a generic pornographic content detector. Their purposed method consists of three part. The first part is "instance generation" where multiple instances of an image are created using a sliding window technique, then resized and divided into multiple segments. The second part is "instance selection" where, using a semi-automated process, they search for private parts of humans. Finally, the third part is the "DCNN-based feature learning" which takes the selected images as input. Similarly, Nian et al. [49] proposed a pornographic image content CNN detector. They retrained the pre-trained ImageNet and fine-tuned it to detect pornographic images of any scale in a single forward pass.

Consumption of pornographic content is socially acceptable amongst adults. However, this becomes a problem when children are exposed to such content. The porn industry has introduced numerous categories and types of sexual activities, no method which further classifies them was encountered. Pornography detection methods have also been utilised for Indecent Images of Children (IIOC), however, the reliability of these methods is questionable as the performance of these methods cannot be publicly tested.

D. TAMPERING DETECTION

Historically, media tampering was computationally expensive and was only under the reach of big graphic studios [52]. This has changed, and now even a novice user can perform tampering on their personal machine with, widely available, specialized software.

Conventional methods of tampering have been replaced with more advanced methods that cannot only tamper but can also generate fake media from scratch using DL (aka, deepfake). Such material is useful to promote disinformation, propaganda, and influence. Social media became a prime vehicle for distribution of such fabricated or tampered content [53], [54]. Table 4 summarises the surveyed DL methods for tampering (and deepfake) detection.

Tampering detection methods can either be blind or non-blind. Birajdar et al. [66] provide a taxonomy of blind techniques for digital image forgery. Some of the tampering operation for images include cloning, retouching, re-sampling,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

TABLE 2. DL solutions for Violence Detection

Method	Model	Media	Year	Dataset Used
FUDAN-NJUST [40]	DNN-SVM	Video	2014	VSD 2014
FUDAN-HUAWEI [41]	LSTM-CNN	Video	2015	VSD 2015
NII-UIT [42]	HOG-MBH-SIFT-MFCC-VDFULL-CNN	Video	2015	VSD 2015
KIT [43]	GIST-IDT-CNN	Video	2015	VSD 2015
MIC-TJU [44]	IDT-SIFT-MFCC-HSH-CNN	Video	2015	VSD 2015
ResnetCrowd [12]	Residual-CNN	Image	2017	Multi Task Crowd [12]
Peixoto et al. [35]	CNN	Image	2018	VSD 2013

TABLE 3. DL solutions for Pornography Detection

Method	Model	Media	Year	Dataset Used
AGNet. [5]	CNN	Image	2015	Pornography-800 and Pornography-2K
SD-MIL [48]	CNN	Image	2016	Pornography-800, Pornography-2K & Unnamed [48]
Nian et al. [49]	CNN	Image	2016	Unnamed [49]
Perez et al. [47]	Two-Stream CNN	Video	2017	Pornography-800 & Pornography-2K
ACORDE. [50]	LSTM-CNN	Video	2018	Pornography-800 & Pornography-2K
Da Silva et al. [51]	Spatio-Temporal CNN	Video	2018	Pornography-800

TABLE 4. DL solutions for Tampering Detection

Method	Model	Media	Year	Dataset Used
Rao et al. [55]	CNN	Image	2016	CASIA & Columbia
Zhang et al. [56]	Auto-encoder	Image	2016	CASIA
Cozzolino et al. [57]	CNN	Image	2017	Unnamed [57]
Bappy et al. [58]	LSTM-CNN	Image	2017	NIST [59], IEEE Forensics Dataset & Coverage
Zhou et al. [60]	Two-Stream Faster R-CNN	Image	2018	NIST [59], CASIA, Coverage & Columbia
ForensicTransfer [61]	CNN Auto-encoder	Image	2018	Face Forensics
Nguyen et al. [62]	Capsule-CNN	Image	2018	DeepFake [63], Reply Attack[2], Face Forensics & Unnamed[3]
Guera et al. [64]	LSTM-CNN	Video	2018	Unnamed [64]
Li et al. [65]	LSTM-CNN	Video	2018	CEW[4] & EBV [65]
MesoNet. [63]	CNN	Video	2018	DeepFake [63] & Face2Face [63]

and copy-move. Previously, videos were difficult to tamper successfully. However, newer deep methods have made it possible to tamper and even generate fake videos [67], [68] that are more resistant to detection [69]. Due to these factors, a renewed interest in tampering detection has also risen, and many DL methods have been proposed.

Rao et al. [55] proposed a 10-layer CNN network, where the first layer is actually a high pass filter. This was carried out in order to generate SRM (Spatial Rich Models) residual maps. The network is strictly designed to detect copy-move and image splicing manipulation. This is an example of a constrained CNN, where the network is forced to learn specific features since SRM features help in detection of image manipulation.

Cozzolino et al. [57] stated that it is not necessary to constrain the CNN, rather a residual-based descriptors can learn specific manipulation operations. They proposed a CNN network that detects image tampering by combining the SRM features with a CNN network. They also generated a synthetic dataset of manipulated images, taken from 4 smartphones and 5 cameras. Cozzolino et al. [61] later proposed an Auto-Encoder Network for image forgery detection. This method was designed in a way that it could quickly adapt to other types of tampering. They also experimented with different variants of their network, where they performed high pass filter on residual images.

Zhang et al. [56] proposed a deep Stacked Auto-Encoder (SAE) network for image manipulation detection. They claimed that their DL model can detect tampering for different image formats. Their network is trained in a two-step manner. The first step learned complex features using the SAE model, while the second step identified the tampered regions by context learning.

The method proposed by Bappy et al. [58] used a combination of LSTM and CNN. This DL method is able to detect multiple manipulation techniques which include: copy-move, image splicing, and removal. Their network first classifies between manipulated vs. non-manipulated images, and then the manipulated parts of the images are highlighted.

Zhou et al. [60] proposed a DL solution to detect tampered images. Their proposed model is composed of a Two-Stream faster R-CNN inspired network, Where the first part is the RGB stream which focuses on the visual cues of the image.

[2]https://dl.gi.de/bitstream/handle/20.500.12116/18295/183.pdf?sequence=1
[3]https://hal-upec-upem.archives-ouvertes.fr/hal-01664590/document
[4]http://parnec.nuaa.edu.cn/xtan/data/ClosedEyeDatabases.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

The second part is the noise stream which focuses on local noise distributions of the image in order to locate the areas where possible tampering might have taken place.

Nguyen et al. [62] developed a DL method which employs capsules for forgery detection in images and videos. Their network is a combination of VGG-19 with capsules. They detect faces and then resize them to images of dimension 128×128. Afterward, these inputs are passed to three primary capsules, which are connected to two capsules which distinguish between real and fake. The attacks on which they focus include Replay Attack, Face Swapping, Facial Re-enactment, and Fully Computer-Generated Image Detection.

Newer DL-based methods are now emerging to generate fake multimedia. Antipov et al. [70] used GAN to produce images for human aging. The goal was to predict how an individual would age over a period of time. Another GAN-based DL method proposed by Huang et al. [71] could generate the image of a frontal face from a given image which has a side view of an individual's face. This method is called Two-Pathway Generative Adversarial Network (TP-GAN).

Many DL methods have been developed to tackle such deep methods that can generate fake media. Guera et al. [64] proposed a convolutional-LSTM network that is specifically designed for detecting deepfake videos. They used a total of 600 videos to train their network which achieved an accuracy of 97.1%. Li et al. [65] also trained a network based on CNN and LSTM. This network took into consideration eye blinking, which is a physiological and behavioural trait of humans, to detect deepfake videos. Two distinct DL networks were proposed by Afchar et al. [63] to detect face tampering in fabricated videos produced using DeepFake and Face2Face softwares individually. According to them no single network could detect tampered video generated by these two softwares. The two networks were named as Meso-4 and MesoInception-4.

DL methods have now been actively utilising as tampering detection techniques. Datasets for deepfakes have recently been produced to tackle this emerging threat that can compromise trust on digital media been shared on the Internet. DL enabled manipulation and fake content generation will be an interesting area of research in the near future.

E. AGE ANALYSIS

Aging is a complex biological phenomenon that manifests differently in every individual and can be affected by external and internal factors such as genetic makeup, disease, drug abuse, habitat, and environment. Automatic age retrieval methods can be categorised into "age classification" and "regression tasks". In the former, individuals are assigned to an age group. In the latter, a numeric value is predicted and this can be further sub-categorised into "Apparent age" and "Biological age estimation". Angulu et al. [10] surveyed age estimation using facial images of people.

Many DL-based methods have been developed for age analysis. Dong et al. [72] proposed a DL solution where they detect the face through five facial key points. This facial image is then passed through a network which outputs an age range based on these facial features. To cope with the lack of a comprehensive and large-scale dataset, they used transfer learning to train the model. Further, they proposed a new loss function. Networks configuration included four convolutional layers, followed by max-pooling layers, and one fully connected layer.

During the Chalearn LAP competition, Antipov et al. [73] proposed a DL network for apparent age estimation securing the first position. Their network is inspired by the VGG-16 for facial recognition. During the span of this competition, they generated the IMDB-Wiki dataset on which they trained their network. They also fine-tuned the network for the precise age estimation of children between 0 and 12 years old.

Xing et al. [7] performed a detailed analysis of DL models and strategies for the problem of age estimation. They studied model formulation, architectures, and selection of loss functions. In doing so, they proposed a multi-task CNN for age, race and gender incorporating all the learned insights. Network variants included a very deep multi-task and hybrid multi-task and hybrid multi-task learning architectures.

A major challenge in age estimation is the variety of races and genders which exhibit different aging patterns. Keeping this in mind, Li et al. [74] proposed a DL solution called Deep Cross-Population (DCP) age estimation model. They presented two novel loss functions: (1) the Cost-Sensitive multitask loss function, and (2) the order-preserving pairwise loss function.

Rothe et al. [75] proposed a VGG-16 [17] inspired architecture for apparent and real age estimation, called DEX. They trained a network with their IMDB-WIKI dataset and secured the first position in the ChaLearn LaP 2015 challenge on apparent age estimation. Inspired by this, Agustsson et al. [14] proposed a new dataset and DL solution for age estimation. Their method is basically a residual DEX. Their residual regression network is designed after studying the relationship between real and apparent age which further improved their performance.

Aging is an uncontrolled and irreversible process of the human body. The most visible effects of aging are exhibited on facial features. These changes are very personal for each individual, however, the age estimation methods perform well at distinguishing between the extremes in the age distribution. The performance is degraded when age groups closer to each other need to be classified.

Table 5 summarises the discussed DL solutions for age analysis.

F. SCENE RECOGNITION

Scene recognition involves the semantic understanding of visual entities that share a common context (objects and background), and this is a difficult task to automate [76].

It is used in many types of application such as content-based indexing and retrieval systems, robotics, crime scene analysis [77], and 3D scene construction [78]. Scenes are

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

TABLE 5. DL solutions for Age Analysis

Method	Model	Media	Year	Dataset Used
DEX [75]	CNN	Image	2015	ChaLearn LAP 2015 & IMDB-Wiki
Dong et al. [72]	CNN	Image	2016	The Images of Groups[5]
Antipov et al. [73]	Multi-CNN	Image	2016	IMDB-Wiki, ChaLearn LAP 2016 & Children [73]
Xing et al. [7]	Multi-Task CNN	Image	2017	Morph-II & WebFaces
Residual DEX [14]	CNN	Image	2017	APPA-REAL
DCP [74]	CNN	Image	2018	Morph-II & WebFace

broadly classified as indoor and outdoor with high intra-class and inter-class variation.

Zhou et al. [79] proposed a novel measure to gauge density and diversity bias; they applied it to different datasets. Through the visualisation of object-centric and scene-centric CNN, they realised that objects and scenes have different internal representations, and concluded that the ImageNet-trained CNN performs worse than the Place-CNN. Based on this knowledge, they proposed a Hybrid-CNN trained on both objects and places, and this approach achieved a performance boost. Herranz et al. [80] improved upon the work of Zhou et al. by removing the scale-induced bias, and combining the object and scene features. According to them, both Places-CNN and ImageNet-CNN were trained on images with different scale ranges which caused performance degradation. To remove this bias, they presented a Multi-Scale architecture with scale-specific networks, which improves recognition accuracy.

Wang et al. [81] combined the traditional and CNN based features extractors. They proposed an end-to-end architecture, called PatchNet, trained in a weakly supervised manner. The features learned by PatchNet were then complemented by a new image representation scheme, called Vector of Semantically Aggregated Descriptors (VSAD). Together, PatchNet and VSAD show superior performance. This is another example where traditional features combined with DL features have improved performance. Another hybrid approach was proposed by Guo et al. [82], called Locally-Supervised Deep Hybrid Model (LS-DHM). They use Local Convolutional Supervision (LCS) layer and Fisher Convolutional Vector (FCV), integrated with the learned feature representation of LS-DHM.

To tackle inter-class similarity and intra-class variation, Kim et al. [83] proposed a hierarchical network, which consists of alternating specialist networks based on a binary tree structure. The specialist and generalist models output the same number of predictions while using both global ordered and orderless pooling architectures delivering better performance than other tree structured networks.

Another method for scene recognition, Adi-Red was proposed by Zhao et al. [84]. This method uses a discriminative discovery network (DisNet) to generate Dis-Maps which provides discriminative regions for the given images. These Dis-Maps are then aggregated within a multi-scale framework. It was claimed that Adi-Red was the first method to use discriminative regions in an adaptive fashion for scene recognition.

Liu et al. [85] proposed a Dictionary Learning Layer (DLL) which is composed of recurrent units. They replaced the fully connected layer and ReLu with the newly designed DLL layer. According to them, DLL layers learn optimal dictionaries enabling the extraction of high discriminative and sparse features. Furthermore, they proposed to deploy some constraints to avoid over-fitting based on the advantages of Mahalanobis and Euclidean distance. They also proposed a new label discriminative regressor. They call their network CNN-DL.

Scene recognition has received much attention in computer vision. Many large scale datasets for scene recognition are available. The biggest challenge that causes performance loss in scene recognition is the inter and intraclass variation. DL methods for scene recognition have produced a state-of-the-art performance on these benchmarks. Table 6 summarises the reviewed DL models for Scene Recognition.

G. PERSON RE-IDENTIFICATION

This class of problem is concerned with the re-identification of a particular individual (previously observed in an image or video) at different non-overlapping views across a period of time, from multiple cameras viewpoints under different poses. Person Re-Identification (Re-ID) is very challenging even for humans [86]. Real-world applications include multi-camera tracking of criminals or individuals of interest, robotics human-machine interactions, crowd traffic analysis, and management [87]. A typical Re-ID system consists of three main components: person detection, person tracking, and person retrieval [88].

Single-shot and multi-shot recognition strategies are used for Re-ID task [89]. A survey by Bedagkar et al. [9] elaborated on trends, methods, datasets, and taxonomy of Re-ID approaches. Advancements in video surveillance have motivated the development of recent DL solutions for Re-ID.

Li et al. [90] proposed a six-layer Filter Pairing Neural Network (FPNN), which jointly optimized the Re-ID pipeline including feature extraction, photometric and geometric transforms, misalignment, occlusions, and classification. They use verification loss function to train their

[5]http://chenlab.ece.cornell.edu/people/Andy/ImagesOfGroups.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

TABLE 6. DL solutions for Scene Recognition

Method	Model	Media	Year	Dataset Used
Zhou et al. [79]	CNN	Image	2014	Places, ImageNet , MIT Indoor67[6], Scene15[7], Sun-205, Sun397, Sun Attribute, Caltect101[8], Caltect256[9], Action40[10] & Event8
Herranz et al. [80]	Multi-Scale CNN	Image	2016	Scenes15[7], MIT Indoor67[6] & SUN397
PatchNet [81]	VASD-CNN	Image	2017	Sun397 & MIT Indoor67[6]
LS-DHM [82]	FCS-LCS-CNN	Image	2017	Sun397 & MIT Indoor67[6]
Kim et al. [83]	Hierarchical-CNN	Image	2018	Sun397, Places205 & CIFAR100[11]
Adi-Red [84]	Multi-Scale CNN	Image	2018	Sun397 & Places365
CNN-DL [85]	CNN	Image	2018	Sun397, MIT Indoor67[6] & Scenes15[7]

network.

Ahmad et al. [91] presented a deep network which in addition to feature representation also learns the similarity metric through a novel layer that calculates the cross-input neighbourhood differences. This layer compares the features of the neighbouring location to capture the local relationship of the images; this approach was followed after two convolution layers in their model. They provided a detailed comparison of their model with other deep architectures which included FPNN [90].

Xiao et al. [20] proposed a DL solution for Re-ID task where they employed a domain guided dropout layer. Their network learns generic features from six domains (i.e., camera views captured by different datasets) to solve the problem at hand. They first pre-trained their network on the combined dataset and claimed that this strategy provides a strong baseline model that can be retrained for individual domains. Afterward, they replaced the standard dropout layer with their own domain guided dropout layer.

Most of the methods benefit from overlapping regions in the images to solve the Re-ID task. However, the DL model by Cheng et al. [92] did not rely on them. Their "Multi-Channel Part Based CNN" model was trained using a triplet loss function that learns global full-body and local body-parts of the person under observation.

Wang et al. [93] proposed a Joint Attribute-Identity DL (TJ-AIDL) that is capable of transferring learned features in an unsupervised manner reducing the need for large scale datasets. Their network consists of two parallel CNN networks followed by an auto-encoder were they use a learning strategy of attributes and identity discrimination. Li et al. [94] also proposed an unsupervised DL algorithm, called Tracklet Association Unsupervised DL (TAUDL) framework. Their method does not require labelled camera pairwise images.

Chen et al. [95] devised a DL solution that works by learning features for scale-specific and multi-scale person appearance, as opposed to most single-scale methods. They proposed a novel deep Pyramid Feature Learning (DPFL) CNN model.

Zheng et al. [96] combined a CNN-based verification and identification model for the Re-ID task. Their siamese network computes and combines the verification and identification loss in order to generate a highly discriminative pedestrian embedding and similarity measure at the same time.

Wu et al. [97] presented a video-based DL solution that uses a stepwise learning method (EUG: Exploit the Unknown Gradually) to enhance the discriminative capability of their model and predict pseudo labels. They use only one labelled tracklet to re-identify the other unlabelled tracklets by employing a progressive sampling strategy for single-shot Re-ID.

Re-ID has emerged as a relatively newer task in computer vision as many new methods and datasets have been proposed. However, Re-ID in open world is still a challenging problem due to multi camera angle and view, lack of universal feature representation. Further, there is a lack of standardised evaluation criteria for Re-ID task. Table 7 compiles the approaches discussed in this section for Re-ID.

H. GAIT RECOGNITION

Gait is a trait or signature, that can be used as biometric or behavioural identifier therefore it can be used to distinguish one individual from others [98]. Such signature may be composed of many factors which include: pattern of the human walk (i.e., length and movement of torso and limbs), weight, arm swing, and musculoskeletal structure of the body. Gait is affected by many external factors which can include footwear, clothing, walk speed, injury, and mood [99], [100]. A gait-based biometric system has certain advantages over others. Firstly, it can be used to identify a person from far distances. Secondly, it does not require the approval of the individual under observation [101]. Third, no physical interaction is required with the biometric systems. UK and Denmark are using gait recognition to convict criminals through evidence collection and forensic identification [102], [103]. Gait recognition algorithms are generally divided into two categories: Model-based and Appearance-based methods [101], [104]. This paper only focuses on model-free gait

[6]http://web.mit.edu/torralba/www/indoor.html
[7]https://figshare.com/articles/15-Scene_Image_Dataset/7007177
[8]http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[9]http://www.vision.caltech.edu/Image_Datasets/Caltech256/
[10]http://vision.stanford.edu/Datasets/40actions.html
[11]https://www.cs.toronto.edu/ kriz/cifar.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

TABLE 7. DL solutions for Person Re-Identification

Method	Model	Media	Year	Dataset Used
FPNN [90]	CNN	Image	2014	CUHK03 & CUHK01
Ahmad et al. [91]	Two-Stream CNN	Image	2015	CUHK03, CUHK01 & VIPeR
Xiao et al. [20]	CNN	Image	2016	CUHK03, CUHK01, PRID, VIPeR, 3DPeS & iLIDs
MCP-CNN [92]	Multi-Channel CNN	Image	2016	iLIDs, VIPeR, PRID2011[12] & CUHK01
Zheng et al. [96]	CNN	Image	2018	CUHK03, Market1501[14] & Oxford5K[13]
TJ-AIDL [93]	Multi-branch CNN	Image	2018	VIPeR, PRID2011[12], Market1501[14] & DukeMTMC-ReID[15]
DPFL [95]	Multi-Channel CNN	Image	2018	Market1501[14], CUHK03 & DukeMTMC-ReID[15]
Wu et al. [97]	CNN	Video	2018	MARS & DukeMTMC-ReID[15]
TAUDL [94]	CNN	Image	2018	CUHK03, Market-1501[14], DukeMTMC-ReID[15], iLIDs, PRID & MARS

recognition methods, as they require visual content analysis.

Gait methods rely on Gait Energy Image (GEI), generated by aggregating the silhouette sequences of the person under observation at the expense of losing temporal information. The motivation for silhouette extraction is that it removes color, clothing and other textures from the image.

Using GEI, Shiraga et al. [105] proposed a rather simple CNN, called GEINet. This is a 4 layer network consisting of 2 convolution layer followed by 2 fully connected layers and a softmax layer at the end. Each convolution layer is followed by a pooling and normalisation layer. They use the cross-entropy loss to train their network.

Wu et al. [106] presented another deep CNN method using GEI for cross-view gait recognition. They performed an extensive empirical evaluation for larger cross-view angles. Their method is robust to changing viewpoints and walking conditions, showing greater generalisation ability across multiple larger datasets.

Castro et al. [107] proposed a CNN method inspired by the Two-Stream Network [23]. Their network takes Optical Flow Maps as input images, rather than GEI, and generates gait probabilities for every individual. Another method which did not use GEI images was proposed by He et al. [108]. They presented a Multi-Task Generative Adversarial Network (MGAN) that learns view specific feature representations. Their network is composed of 5 components: Encoder, View-angle classifier, View transfer layer, Generator, and Discriminator. They also utilised Period Energy Image (PEI), which is a multi-channel gait template.

Chao et al. [109] proposed Gaitset – an end-to-end DL model that uses "Set Pooling" operations to aggregate silhouette frame-level features. These features are then mapped to a higher discriminative space using Horizontal Pyramid Mapping.

Gait methods heavily rely on GEI images, all methods use them to discard unwanted visual artefacts for gait recognition. Developing robust gait systems is still a challenging task as there are many factors that can affect the performance of these systems such as camera view, clothing, shoe type or carrying objects. Further, if the observed individual is aware of gait systems they can intentionally change their gait. This new research suffers from lack of new gait datasets which are suitable for DL based methods. Table 8 summarises the reviewed DL models for Gait Recognition.

[12]https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/prid11/

[13]http://www.robots.ox.ac.uk/ vgg/data/oxbuildings/

[14]http://www.liangzheng.com.cn/Datasets.html

[15]https://megapixels.cc/datasets/duke_mtmc/

IV. DATASETS USEFUL FOR DL SOLUTIONS TO ADDRESS VISUAL CONTENT ANALYSIS PROBLEMS

DL is data-hungry in nature [110], thus, the availability of large scale, high quality, publicly available datasets play a significant role in attracting the research community. This section provides a list of benchmark datasets for the problems discussed in Section III. Table 9 lists the dataset name, the target domain area, number of training examples, media type, and publication year. This table was compiled based on most authoritative, recently published and tested benchmarks; it is not meant to be comprehensive.

Action recognition, violence and pornography detection datasets predominantly target videos. AcitivtyNet [111], Hollywood2 [112] and UCF-101 [113] are the most cited benchmarks for action recognition, but SLAC [114] has the highest number of examples. In contrast to action recognition, fewer datasets are available for violence detection. VSD [34] is a freely available dataset for content based violence detection. However, the largest violence dataset is BEHAVE [115].

Pornography detection datasets are bound by ethical concerns and, therefore, are scarce. The biggest datasets available are Pornography-800 [116] and Pornography-2000 [117] containing 800 and 2000 videos, respectively.

A renewed interest has developed in tampering detection due to the recent proliferation of fake media, and deepfakes. NIST has been hosting a series of Media Forensics Challenges since 2016, specifically designed to facilitate the development of tampering detection methods. MFC-2018 [59] was the last dataset released at time of writing through this competition. Traditional datasets include the Columbia dataset [118] and CASIA [119]. Coverage [120] is another dataset which only contains around 100 authentic-tampered image pairs which were generated after performing 6 tampering operations. Rossler et al [121] have produced

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

TABLE 8. DL solutions for Gait Recognition

Method	Model	Media	Year	Dataset Used
GEINet [105]	CNN	GEI-Image	2016	OU-ISIR[16]
Castro et al. [107]	CNN	OF-Image	2017	TUM-GAID[17]
Wu et al. [106]	CNN	GEI-Image	2017	CASIA-B, OU-ISIR[16], USF[18]
Gaitset [109]	CNN	Silhouette-Image	2018	CASIA-B and OU-ISIR (MVLP)[19]
MGAN [108]	GAN	PEI-Image	2019	OU-ISIR[16], CASIA-B & USF[18]

a large scale dataset, Face Forensics, which consists of 1004 videos applying two types of manipulation: source-to-target and self-re-enactment.

All age analysis datasets consist of facial images. MORPH [122] and FG-net [123] are the oldest and most highly cited datasets. IMDB-WIKI [75] and CACD [124] are larger dataset and they have been published more recently. FG-NET [123], MORPH [122], AdienceFaces [125], CACD [124], IMDB-WIKI [75] and AgeDB [126] contain exact ages as labels, whereas in Gallagher [127], VADANA [128] and AdienceFaces [125] the images are assigned to age groups.

Specific datasets for scene recognition include: SUN [129] (the most cited), MS-COCO [130] and Places [131] (the most recent), and TinyImage [132] (the largest, containing around 80 million images).

Most of the Person Re-Identification datasets are captured as videos, then bounding boxes are drawn to allow them to be used for training Re-ID methods. They are compiled under different circumstances such as number of cameras, number of identities/individuals, single-shot or multi-shot. The most recent datasets are RPIfield [133], MSMT17 [134] and the Motion Analysis and Re-identification Set (MARS) [135]. MSMT17 has 126,441 bounding boxes. MARS contains around 20,000 video sequences producing a real-world large scale dataset; it is an extension of the Market-150 [136], which is an image based benchmark for person re-identification. The mostly cited image-based benchmark is VIPeR [137]; it contains 1,264 images with 632 identities. Traditional gait recognition datasets only had single view images; CASIA-A [138] and CMU Mobo [139] are amongst the oldest gait benchmarks for images. In terms of video gait datasets, CASIA-B [140] is the largest.

V. SHORTCOMINGS OF DEEP LEARNING

In Sections I to III, we have praised DL's advantages which turn it into a promising, and increasingly explored solutions for visual content analysis tasks. One strong advantage is its ability to learn features without the need for pre-defined expert knowledge informed, feature extractors.

However, deep methods have some major limitations. For example, in order to extract features, the very first pre-condition that needs to be fulfilled is the presence of high quality and high volume data [131], [162]. As DL is data-hungry for training in nature [110], it totally depends on the dataset used for learning feature representation. This can negatively affect the ability to generalise results. Therefore, the presence of any type of bias in the dataset [163] – in terms of capture, selection, negative set, and example variety – will compromise the quality of data and, consequentially, the output quality of the DL model. Very authoritative, large and reliable benchmarks are prerequisites for training robust deep networks.

DL methods are "black boxes" in nature [164], [165]. Efforts have been made to better understand the learning process of a given network through output visualisation of layers, e.g., by Zeiler et al. [16]. DL networks are self-contained, therefore, debugging them is not yet possible [166]. Unaware of the internal working if DL methods are employed in safety-critical systems, they can be a possible avenue for sabotage or simply malfunction. This can turn them into a security concern. When applied to the different domains discussed in Section III.

Concerns have been previously raised in relation to convolution operation, which is the highly used operation in deep methods for visual tasks. This operation does not pay attention to pose, texture and deformations [167]. In addition, just applying convolution operation is not enough for performance gains in many content analysis problems. Traditional handcrafted features for motion or sounds (such as STIP, IDT and MFCC) have increasingly been combined with deep networks, which compromises the end-to-end nature of the model [43], [44].

A. RESEARCH DIRECTIONS

DL is now been deployed in time and safety-critical systems. The black box nature of these methods raise privacy and security concerns, creating a trust deficit, and causing reluctance to fully rely on them. The development of debugging technologies would greatly help in the understanding of the learning mechanism of Deep Networks [21]. This would further contribute towards the development of new architectures, methods and provide opportunities for the optimisation of existing methods.

For visual content analysis on videos, the survey uncovered that methods just based on CNN are ineffective and are often combined with other motion-based feature extractors. This is derived from the fact that temporal and motion

[16]http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitLP.html
[17]https://www.mmk.ei.tum.de/en/misc/tum-gaid-database/
[18]http://www.eng.usf.edu/cvprg/Gait_Data.html
[19]http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitMVLP.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924733, IEEE Access

TABLE 9. List of datasets for content analysis of multimedia problems

No	Dataset Name	Category	Instances	Media	Year
1	Hollywood 2 [112]	Action Recognition	1,694	Video	2009
2	UFC-101 [113]	Action Recognition	13,320	Video	2012
3	ASLAN [141]	Action Recognition	3,697	Video	2012
4	HMDB51 [142]	Action Recognition	6,766	Video	2011
5	ActivityNet-200 [111]	Action Recognition	28,108	Video	2015
6	DALY [143]	Action Recognition	8,133	Video	2016
7	Kinetics [144]	Action Recognition	306,245	Video	2017
8	20BN-something-something [145]	Action Recognition	108,499	Video	2017
9	SLAC [114]	Action Recognition	1,750,000	Video	2017
10	VLOG [146]	Action Recognition	114,000	Video	2017
11	Moments In Time [147]	Action Recognition	1,000,000	Video	2018
12	Epic kitchen [148]	Action Recognition	39,596	Video	2018
13	BEHAVE [115]	Violence Detection	83,545	Video	2010
14	Crowd Violence: Non-violence Database and benchmark [38]	Violence Detection	246	Video	2012
15	National Hockey league and Movies [149]	Violence Detection	1,000	Video	2011
16	Violent Scene Dataset [34]	Violence Detection	32,678	Video	2015
17	Pornography-800 [116]	Pornography Detection	800	Video	2013
18	Pornography-2k [117]	Pornography Detection	2,000	Video	2016
19	Columbia dataset [118]	Tampering Detection	2,208	Image	2004
20	IEEE Image Forensics Challenge Dataset [150]	Tampering Detection	2200	Image	2,013
21	CASIA [119]	Tampering Detection	14,044	Image	2013
22	Media Forensics Challenge 2018 (MFC2018) [59]	Tampering Detection	5,000,000	Image	2016
23	Coverage [120]	Tampering Detection	200	Image	2016
24	Face Forensics [121]	Tampering Detection	1,004	Video	2018
25	FG-net [123]	Age Analysis	1,002	Image	2002
26	MORPH [122]	Age Analysis	55,134	Image	2006
27	Gallagher [127]	Age Analysis	28,231	Image	2009
28	VADANA [128]	Age Analysis	2,298	Image	2011
29	Cross Age celebrity dataset [124]	Age Analysis	163,446	Image	2014
30	AdienceFaces [125]	Age Analysis	26,580	Image	2014
31	Chalearn dataset for apparent age estimation [151]	Age Analysis	4,691	Image	2015
32	IMDB-WIKI dataset [75]	Age Analysis	524,230	Image	2015
33	AgeDB [126]	Age Analysis	16,488	Image	2017
34	APPA-REAL [14]	Age Analysis	7,591	Image	2017
35	TinyImage [132]	Scene Recognition	79,302,017	Image	2008
36	SUN database [129]	Scene Recognition	899	Image	2010
37	MS-COCO [130]	Scene Recognition	2,500,000	Image	2014
38	Places [131]	Scene Recognition	10,000,000	Image	2018
39	VIPeR [137]	Person Re-Identification	1,264	Image	2007
40	3DPES [152]	Person Re-Identification	1,000	Video	2011
41	CUHK01 [153]	Person Re-Identification	1,942	Image	2012
42	CUHK02 [154]	Person Re-Identification	7,264	Image	2013
43	CUHK03 [90]	Person Re-Identification	13,164	Image	2014
44	iLIDS-VID [155]	Person Re-Identification	600	Video	2014
45	Market-150 [136]	Person Re-Identification	32,668	Image	2015
46	MARS [135]	Person Re-Identification	20,715	Video	2016
47	MSMT17 [134]	Person Re-Identification	126,441	Video	2018
48	RPIfield [133]	Person Re-Identification	601,581	Video	2018
49	CASIA-A [138]	Gait Recognition	19,139	Image	2002
50	CMU Mobo [139]	Gait Recognition	204,000	Image	2004
51	CASIA-B [140]	Gait Recognition	13,640	Video	2006
52	CASIA-C [156]	Gait Recognition	1,530	Video	2006
53	Southampton Dataset [157]	Gait Recognition	600	Image	2006
54	TokyoTech [158]	Gait Recognition	1,902	Video	2010
55	Soton multimodal [159]	Gait Recognition	1,986	Video	2011
56	AVA [160]	Gait Recognition	1,200	Video	2014
57	KY4D [161]	Gait Recognition	672	Video	2014

information is of vital importance in video content analysis. To tackle this, a 3D convolution layer may be used to process the extra dimension when training end-to-end models [27].

Another research development is the use of capsule network-based methods, as proposed by Nguyen et al. [62]. Capsule networks are designed to remove the limitation of convolution networks inability to utilise spatial hierarchical information and its relationship with orientation during feature extraction. Capsules incorporate the viewpoint changes through voting for pose matrix, trained using "routing-by-agreement" algorithm. which only passes features to the higher level capsules if an agreement is reached. Due to this, they require far fewer data to train compared to CNN. These advantages make them a very attractive option for the future development of DL methods.

Development of sophisticated techniques such as DeepFake and Face2Face has made it very easy to tamper and develop digital media. DL techniques are now able to translate human emotions from a source person to a target video. Due to these alarming developments, in near future DL techniques would also be utilised for deepfake detection. As these DL methods are able to generate fake media in real time. There is a possibility that traditional tampering detection techniques would fail against these deepfake methods opening up new research gaps.

Even though action recognition methods have received great attention from the research community, violence detection in comparison is overlooked. The fact that there is a lack of comprehensive dataset for violence detection seriously hinders the development of robust DL methods. Incidents of real-life violence are very different from the ones presented in movies. Violence being a subjective matter makes the development for an extensive dataset difficult. However. we believe that violence detection methods have huge application potential especially in the world where smart cities are becoming a reality.

VI. CONCLUSION

This paper identified eight problems related to visual content analysis. For each class of problem, we reviewed the state-of-the-art and acknowledged the best performing DL methods proposed in the literature. We also provided a compilation of authoritative datasets useful for training deep methods to address those problem domains. Finally, we discussed the potential limitations of DL methods that can negatively affect their reliability, robustness, and accuracy for visual content analysis.

The survey adopted a breadth-first strategy rather than a deep-first strategy. This means that we aimed at covering DL solutions for all eight problems in detriment of providing a very detailed account of individual problem classes. When applicable, we pointed to other surveys dedicated to specific problems. The rationale for this approach was the cross-fertilisation of DL methods which can potentially be re-applied from one visual problem analysis to another.

We found that violence detection was the most overlooked content analysis problem among the eight classes surveyed. The root cause seems to be the fact that it is considered as a subcategory of action recognition, hindering its development. In contrast, the Person Re-Identification problem gained momentum recently with many newly developed deep methods proposed in the literature. Detection of deepfakes is also becoming an alarming challenge, as DL-based tampering methods are becoming more sophisticated. Capsule networks are emerging as the new feature extractors that have the potential to replace convolution layer in the DL methods of future.

• • •

REFERENCES

[1] E. Lee, "How Long Does it Take to Shoot 1 Trillion Photos?" [Online] http://blog.infotrends.com/how-long-does-it-take-to-shoot-1-trillion-photos/ (Last accessed 12/03/2019), 2016.

[2] "Hours of video uploaded to YouTube every minute as of July 2015," [Online] https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/ (Last accessed 12/03/2019), 2015.

[3] R. Carr, "Surveillance politics and local government: A national survey of federal funding for cctv in australia," Security Journal, vol. 29, no. 4, pp. 683–709, 2016.

[4] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," Neurocomputing, vol. 187, pp. 27–48, 2016.

[5] M. Moustafa, "Applying deep learning to classify pornographic images and videos," arXiv preprint arXiv:1511.08899, 2015.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, p. 436, 2015.

[7] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling, "Diagnosing deep learning models for high accuracy age estimation from a single image," Pattern Recognition, vol. 66, pp. 106–116, 2017.

[8] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," Image and vision computing, vol. 60, pp. 4–21, 2017.

[9] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," Image and Vision Computing, vol. 32, no. 4, pp. 270–286, 2014.

[10] R. Angulu, J. R. Tapamo, and A. O. Adewumi, "Age estimation via face images: a survey," EURASIP Journal on Image and Video Processing, vol. 2018, no. 1, p. 42, 2018.

[11] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "A network of deep neural networks for distant speech recognition," in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 4880–4884.

[12] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in 14th International Conf. on Advanced Video and Signal Based Surveillance. IEEE, 2017, pp. 1–7.

[13] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in Advances in neural information processing systems, 2016, pp. 3468–3476.

[14] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe, "Apparent and real age estimation in still images with deep residual regressors on appa-real database," in 12th International Conf. on Automatic Face & Gesture Recognition. IEEE, 2017, pp. 87–94.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[16] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in Proc. of the European Conf. on Computer Vision. Springer, 2014, pp. 818–833.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. of the IEEE Conf. on computer vision and pattern recognition, 2016, pp. 770–778.

[20] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.

[21] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS), vol. 5, 2015, pp. 1–6.

[22] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," arXiv preprint arXiv:1806.11230, 2018.

[23] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems, 2014, pp. 568–576.

[24] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.

[25] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in Proc. of the IEEE International Conf. on Computer Vision, 2015, pp. 4597–4605.

[26] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in Proc. of the IEEE Conf. on computer vision and pattern recognition, 2015, pp. 4694–4702.

[27] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proc. of the IEEE international Conf. on computer vision, 2015, pp. 4489–4497.

[28] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in International Conf. on machine learning, 2015, pp. 843–852.

[29] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in Proc. of the IEEE Conf. on computer vision and pattern recognition, 2015, pp. 2625–2634.

[30] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 1933–1941.

[31] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 3034–3042.

[32] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 6, pp. 1510–1517, 2018.

[33] H. Wang and C. Schmid, "Action recognition with improved trajectories," in Proc. of the IEEE international Conf. on computer vision, 2013, pp. 3551–3558.

[34] C.-H. Demarty, C. Penet, M. Soleymani, and G. Gravier, "Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation," Multimedia Tools and Applications, vol. 74, no. 17, pp. 7379–7404, 2015.

[35] B. M. Peixoto, S. Avila, Z. Dias, and A. Rocha, "Breaking down violence: A deep-learning strategy to model and classify violence in videos," in Proc. of the 13th International Conf. on Availability, Reliability and Security. ACM, 2018, p. 50.

[36] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Temporal robust features for violence detection," in Winter Conf. on Applications of Computer Vision. IEEE, 2017, pp. 391–399.

[37] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," Image and vision computing, vol. 48, pp. 37–41, 2016.

[38] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in Computer Society Conf. on Computer Vision and Pattern Recognition Workshops. IEEE, 2012, pp. 1–6.

[39] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast violence detection in video," in International Conf. on Computer Vision Theory and Applications, vol. 2. IEEE, 2014, pp. 478–485.

[40] Q. Dai, Z. Wu, Y.-G. Jiang, X. Xue, and J. Tang, "Fudan-njust at mediaeval 2014: Violent scenes detection using deep neural networks." in MediaEval, 2014.

[41] Q. Dai, R.-W. Zhao, Z. Wu, X. Wang, Z. Gu, W. Wu, and Y.-G. Jiang, "Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning." in MediaEval, 2015.

[42] V. Lam, S. P. Le, D.-D. Le, S. Satoh, and D. A. Duong, "Nii-uit at mediaeval 2015 affective impact of movies task." in MediaEval, 2015.

[43] P. Marin Vlastelica, S. Hayrapetyan, M. Tapaswi, and R. Stiefelhagen, "Kit at mediaeval 2015-evaluating visual cues for affective impact of movies task." in MediaEval, 2015.

[44] Y. Yi, H. Wang, B. Zhang, and J. Yu, "Mic-tju in mediaeval 2015 affective impact of movies task." in MediaEval, 2015.

[45] S. C. Simon and T. Greitemeyer, "The impact of immersion on the perception of pornography: A virtual reality study," Computers in Human Behavior, vol. 93, pp. 141–148, 2019.

[46] A. Gangwar, E. Fidalgo, E. Alegre, and V. González-Castro, "Pornography and child sexual abuse detection in image and video: A comparative evaluation," 2017.

[47] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Video pornography detection through deep learning techniques and motion information," Neurocomputing, vol. 230, pp. 279–293, 2017.

[48] Y. Wang, X. Jin, and X. Tan, "Pornographic image recognition by strongly-supervised deep multiple instance learning," in International Conf. on Image Processing. IEEE, 2016, pp. 4418–4422.

[49] F. Nian, T. Li, Y. Wang, M. Xu, and J. Wu, "Pornographic image detection utilizing deep convolutional neural networks," Neurocomputing, vol. 210, pp. 283–293, 2016.

[50] J. Wehrmann, G. S. Simões, R. C. Barros, and V. F. Cavalcante, "Adult content detection in videos with convolutional and recurrent neural networks," Neurocomputing, vol. 272, pp. 432–438, 2018.

[51] M. V. da Silva and A. N. Marana, "Spatiotemporal cnns for pornography detection in videos," arXiv preprint arXiv:1810.10519, 2018.

[52] T.-T. Ng and S.-F. Chang, "Identifying and prefiltering images," IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 49–58, 2009.

[53] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–36, 2017.

[54] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy," in Proc. of the 22nd international Conf. on World Wide Web. ACM, 2013, pp. 729–736.

[55] Y. Rao and J. Ni, "A deep learning approach to detection of splicing and copy-move forgeries in images," in International Workshop on Information Forensics and Security. IEEE, 2016, pp. 1–6.

[56] Y. Zhang, J. Goh, L. L. Win, and V. L. Thing, "Image region forgery detection: A deep learning approach." in SG-CRC, 2016, pp. 1–11.

[57] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection," in Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security. ACM, 2017, pp. 159–164.

[58] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. Manjunath, "Exploiting spatial structure for localizing manipulated image regions," in Proc. of the IEEE international Conf. on computer vision, 2017, pp. 4970–4979.

[59] N. Nimble, "Datasets," 2018.

[60] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," arXiv preprint arXiv:1805.04953, 2018.

[61] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," arXiv preprint arXiv:1812.02510, 2018.

[62] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," arXiv preprint arXiv:1810.11215, 2018.

[63] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in International Workshop on Information Forensics and Security. IEEE, 2018, pp. 1–7.

[64] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in 15th International Conf. on Advanced Video and Signal Based Surveillance. IEEE, 2018, pp. 1–6.

[65] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in International Workshop on Information Forensics and Security. IEEE, 2018, pp. 1–7.

[66] G. K. Birajdar and V. H. Mankar, "Digital image forgery detection using passive techniques: A survey," Digital investigation, vol. 10, no. 3, pp. 226–245, 2013.

[67] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," ACM Transactions on Graphics, vol. 36, no. 4, p. 95, 2017.

[68] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, "Deep video portraits," arXiv preprint arXiv:1805.11714, 2018.

[69] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," 2018.

[70] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in International Conf. on Image Processing. IEEE, 2017, pp. 2089–2093.

[71] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in Proc. of the IEEE International Conf. on Computer Vision, 2017, pp. 2439–2448.

[72] Y. Dong, Y. Liu, and S. Lian, "Automatic age estimation based on deep learning algorithm," Neurocomputing, vol. 187, pp. 4–10, 2016.

[73] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Apparent age estimation from face images combining general and children-specialized deep learning models," in Proc. of the IEEE Conf. on computer vision and pattern recognition workshops, 2016, pp. 96–104.

[74] K. Li, J. Xing, C. Su, W. Hu, Y. Zhang, and S. Maybank, "Deep cost-sensitive and order-preserving feature learning for cross-population age estimation," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 399–408.

[75] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in Proc. of the IEEE International Conf. on Computer Vision Workshops, 2015, pp. 10–15.

[76] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proc. of the IEEE Conf. on computer vision and pattern recognition, 2015, pp. 3128–3137.

[77] R. M. Gardner, "A qualitative theory for crime scene analysis," J Assoc Crime Scene Reconstr, vol. 20, pp. 45–55, 2016.

[78] E. Bostanci, "3d reconstruction of crime scenes and design considerations for an interactive investigation tool," arXiv preprint arXiv:1512.03156, 2015.

[79] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in Advances in neural information processing systems, 2014, pp. 487–495.

[80] L. Herranz, S. Jiang, and X. Li, "Scene recognition with cnns: objects, scales and dataset bias," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 571–579.

[81] Z. Wang, L. Wang, Y. Wang, B. Zhang, and Y. Qiao, "Weakly supervised patchnets: Describing and aggregating local patches for scene recognition," IEEE Transactions on Image Processing, vol. 26, no. 4, pp. 2028–2041, 2017.

[82] S. Guo, W. Huang, L. Wang, and Y. Qiao, "Locally supervised deep hybrid model for scene recognition," IEEE transactions on image processing, vol. 26, no. 2, pp. 808–820, 2017.

[83] H. J. Kim and J.-M. Frahm, "Hierarchy of alternating specialists for scene recognition," in European Conf. on Computer Vision. Springer, 2018, pp. 471–488.

[84] Z. Zhao and M. Larson, "From volcano to toyshop: Adaptive discriminative region discovery for scene recognition," arXiv preprint arXiv:1807.08624, 2018.

[85] Y. Liu, Q. Chen, W. Chen, and I. Wassell, "Dictionary learning inspired deep network for scene recognition," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[86] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in Person re-identification. Springer, 2014, pp. 247–267.

[87] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in Person re-identification. Springer, 2014, pp. 1–20.

[88] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," arXiv preprint arXiv:1610.02984, 2016.

[89] T. D'Orazio and G. Cicirelli, "People re-identification and tracking from multiple cameras: A review," in 19th International Conf. on Image Processing. IEEE, 2012, pp. 1601–1604.

[90] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2014, pp. 152–159.

[91] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.

[92] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.

[93] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," arXiv preprint arXiv:1803.09786, 2018.

[94] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in European Conf. on Computer Vision. Springer, 2018, pp. 772–788.

[95] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2590–2600.

[96] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 14, no. 1, p. 13, 2018.

[97] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 5177–5186.

[98] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," Pattern Recognition Letters, vol. 79, pp. 80–105, 2016.

[99] L. Lee and W. E. L. Grimson, "Gait analysis for recognition and classification," in Proc. of Fifth International Conf. on Automatic Face Gesture Recognition. IEEE, 2002, pp. 155–162.

[100] I. Bouchrika and M. S. Nixon, "Exploratory factor analysis of gait recognition," in 8th International Conf. on Automatic Face & Gesture Recognition. IEEE, 2008, pp. 1–6.

[101] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, "Gait energy volumes and frontal gait recognition using depth images," in International Joint Conf. on Biometrics. IEEE, 2011, pp. 1–6.

[102] P. K. Larsen, E. B. Simonsen, and N. Lynnerup, "Gait analysis in forensic medicine," Journal of forensic sciences, vol. 53, no. 5, pp. 1149–1153, 2008.

[103] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon, "On using gait in forensic biometrics," Journal of forensic sciences, vol. 56, no. 4, pp. 882–889, 2011.

[104] M. Alotaibi and A. Mahmood, "Improved gait recognition based on specialized deep convolutional neural network," Computer Vision and Image Understanding, vol. 164, pp. 103–110, 2017.

[105] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in international Conf. on biometrics. IEEE, 2016, pp. 1–8.

[106] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep cnns," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 2, pp. 209–226, 2017.

[107] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. P. de la Blanca, "Automatic learning of gait signatures for people identification," in International Work-Conf. on Artificial Neural Networks. Springer, 2017, pp. 257–270.

[108] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task gans for view-specific feature learning in gait recognition," IEEE Transactions on Information Forensics and Security, vol. 14, no. 1, pp. 102–113, 2019.

[109] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," arXiv preprint arXiv:1811.06186, 2018.

[110] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," Science, vol. 350, no. 6266, pp. 1332–1338, 2015.

[111] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding,"

in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 961–970.

[112] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in Conf. on Computer Vision & Pattern Recognition. IEEE Computer Society, 2009, pp. 2929–2936.

[113] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[114] H. Zhao, Z. Yan, H. Wang, L. Torresani, and A. Torralba, "Slac: A sparsely labeled dataset for action classification and localization," arXiv preprint arXiv:1712.09374, 2017.

[115] S. Blunsden and R. Fisher, "The behave video dataset: ground truthed video for multi-person behavior classification," Annals of the BMVA, vol. 4, no. 1-12, p. 4, 2010.

[116] S. Avila, N. Thome, M. Cord, E. Valle, and A. D. A. AraúJo, "Pooling in image representation: The visual codeword point of view," Computer Vision and Image Understanding, vol. 117, no. 5, pp. 453–465, 2013.

[117] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Pornography classification: The hidden clues in video space–time," Forensic science international, vol. 268, pp. 46–61, 2016.

[118] T.-T. Ng, S.-F. Chang, and Q. Sun, "A data set of authentic and spliced image blocks," Columbia University, ADVENT Technical Report, pp. 203–2004, 2004.

[119] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in China Summit and International Conf. on Signal and Information Processing. IEEE, 2013, pp. 422–426.

[120] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler, "CoverageâĂŤa novel database for copy-move forgery detection," in International Conf. on Image Processing. IEEE, 2016, pp. 161–165.

[121] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," arXiv preprint arXiv:1803.09179, 2018.

[122] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in 7th International Conf. on Automatic Face and Gesture Recognition. IEEE, 2006, pp. 341–345.

[123] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 442–455, 2002.

[124] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in European Conf. on computer vision. Springer, 2014, pp. 768–783.

[125] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," IEEE Transactions on Information Forensics and Security, vol. 9, no. 12, pp. 2170–2179, 2014.

[126] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshop, vol. 2, no. 3, 2017, p. 5.

[127] A. C. Gallagher and T. Chen, "Understanding images of groups of people," in Conf. on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 256–263.

[128] G. Somanath, M. Rohith, and C. Kambhamettu, "Vadana: A dense dataset for facial image analysis," in international Conf. on computer vision workshops. IEEE, 2011, pp. 2175–2182.

[129] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in Computer Society Conf. on Computer Vision and Pattern Recognition. IEEE, 2010, pp. 3485–3492.

[130] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European Conf. on computer vision. Springer, 2014, pp. 740–755.

[131] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 6, pp. 1452–1464, 2018.

[132] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 30, no. 11, pp. 1958–1970, 2008.

[133] M. Zheng, S. Karanam, and R. J. Radke, "Rpifield: A new dataset for temporally evaluating person re-identification," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2018, pp. 1893–1895.

[134] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 79–88.

[135] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in European Conf. on Computer Vision. Springer, 2016, pp. 868–884.

[136] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in Proc. of the IEEE International Conf. on Computer Vision, 2015, pp. 1116–1124.

[137] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3, no. 5. Citeseer, 2007, pp. 1–7.

[138] L. Wang, H. Ning, W. Hu, and T. Tan, "Gait recognition based on procrustes shape analysis." in Proceedings. International Conference on Image Processing, vol. 3. IEEE, 2002, pp. 433–436.

[139] R. Gross and J. Shi, "The cmu motion of body (mobo) database," 2001.

[140] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in 18th International Conf. on Pattern Recognition, vol. 4. IEEE, 2006, pp. 441–444.

[141] O. Kliper-Gross, T. Hassner, and L. Wolf, "The action similarity labeling challenge," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pp. 615–621, 2012.

[142] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in 2011 International Conference on Computer Vision, Nov 2011, pp. 2556–2563.

[143] P. Weinzaepfel, X. Martin, and C. Schmid, "Human action localization with sparse spatial supervision," arXiv preprint arXiv:1605.05197, 2016.

[144] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.

[145] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The something something video database for learning and evaluating visual common sense," in 2017 IEEE International Conference on Computer Vision (ICCV), Oct 2017, pp. 5843–5851.

[146] D. F. Fouhey, W.-c. Kuo, A. A. Efros, and J. Malik, "From lifestyle vlogs to everyday interactions," arXiv preprint arXiv:1712.02310, 2017.

[147] M. Monfort, B. Zhou, S. A. Bargal, A. Andonian, T. Yan, K. Ramakrishnan, L. Brown, Q. Fan, D. Gutfruend, C. Vondrick et al., "Moments in time dataset: one million videos for event understanding," arXiv preprint arXiv:1801.03150, 2018.

[148] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price et al., "Scaling egocentric vision: The epic-kitchens dataset," arXiv preprint arXiv:1804.02748, 2018.

[149] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in International Conf. on Computer analysis of images and patterns. Springer, 2011, pp. 332–339.

[150] I. I.-T. I. F. Challenge, "Datasets," 2013.

[151] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in Proc. of the IEEE International Conf. on Computer Vision Workshops, 2015, pp. 1–9.

[152] D. Baltieri, R. Vezzani, and R. Cucchiara, "3dpes: 3d people dataset for surveillance and forensics," in Proc. of the 2011 joint ACM workshop on Human gesture and behavior understanding. ACM, 2011, pp. 59–64.

[153] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in Asian Conf. on Computer Vision. Springer, 2012, pp. 31–44.

[154] W. Li and X. Wang, "Locally aligned feature transforms across views," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2013, pp. 3594–3601.

[155] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in European Conf. on Computer Vision. Springer, 2014, pp. 688–703.

[156] D. Tan, K. Huang, S. Yu, and T. Tan, "Efficient night gait recognition based on template matching," in 18th International Conf. on Pattern Recognition, vol. 3. IEEE, 2006, pp. 1000–1003.

[157] J. D. Shutler, M. G. Grant, M. S. Nixon, and J. N. Carter, "On a large sequence-based human gait database," in Applications and Science in Soft Computing. Springer, 2004, pp. 339–346.
[158] M. R. Aqmar, K. Shinoda, and S. Furui, "Robust gait recognition against speed variation," in 20th International Conf. on Pattern Recognition. IEEE, 2010, pp. 2190–2193.
[159] S. Samangooei, J. Bustard, M. Nixon, and J. Carter, "On acquisition and analysis of a dataset comprising of gait, ear and semantic data," Multibiometrics for Human Identification, pp. 277–301, 2011.
[160] D. López-Fernández, F. J. Madrid-Cuevas, Á. Carmona-Poyato, M. J. Marín-Jiménez, and R. Muñoz-Salinas, "The ava multi-view dataset for gait recognition," in International Workshop on Activity Monitoring by Multiple Distributed Sensing. Springer, 2014, pp. 26–39.
[161] Y. Iwashita, K. Ogawara, and R. Kurazume, "Identification of people walking along curved trajectories," Pattern Recognition Letters, vol. 48, pp. 60–69, 2014.
[162] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," arXiv preprint arXiv:1506.03365, 2015.
[163] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, "A deeper look at dataset bias," in Domain Adaptation in Computer Vision Applications. Springer, 2017, pp. 37–55.
[164] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," Knowledge-based systems, vol. 8, no. 6, pp. 373–389, 1995.
[165] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," arXiv preprint arXiv:1708.08296, 2017.
[166] G. Ras, M. van Gerven, and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," in Explainable and Interpretable Models in Computer Vision and Machine Learning. Springer, 2018, pp. 19–36.
[167] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in Advances in neural information processing systems, 2017, pp. 3856–3866.

MUHAMMAD SHAHROZ NADEEM received M.Sc. in Computer science from National University of Computer and Emerging Sciences (NUCES), in 2017. He is currently pursuing the Ph.D. in Computer Science at the College of Engineering and Technology, University of Derby, United Kingdom. He also worked with the Department of Computer Science, NUCES as a lecturer from 2017 to 2018. He was also part of the Reveal.ai (Recognition, Vision & Learning) Lab from 2014 to 2017, where he worked on deep learning based image restoration problems. His research interests are data science, image restoration, computer vision, and deep learning.

VIRGINIA N. L. FRANQUEIRA (M'12) received a Ph.D. in Computer Science (focused on Security) from the University of Twente (Netherlands) in 2009, and a M.Sc. in Computer Science (focused on Optimization) from the Federal University of Espirito Santo (Brazil). Since June 2014, she holds a senior lecturer position in Computer Security and Digital Forensics at the University of Derby, UK. She has around 40 publications related to Security or Digital Forensics. She is a member of the British Computer Society and fellow of The Higher Education Academy.

XIAOJUN ZHAI (M'19) received the Ph.D. degree from the University of Hertfordshire, U.K., in 2013. He is currently a Lecturer in the Embedded Intelligent Systems Laboratory at the University of Essex. He has authored/co-authored over 50 scientific papers in international journals and conference proceedings. His research interests mainly include the design and implementation of the digital image and signal processing algorithms, custom computing using FPGAs, embedded systems and hardware/software co-design. He is a BCS, IEEE member and HEA Fellow.

FATIH KURUGOLLU (M'02–SM'08) received the B.Sc., M.Sc., and Ph.D. degrees from Istanbul Technical University, Istanbul, Turkey, in 1989, 1994, and 2000, respectively, all in computer engineering. From 1991 to 2000, he was a Research Fellow with the Marmara Research Centre, Kocaeli, Turkey. In 2000, he joined the School of Computer Science, Queen's University Belfast, Belfast, U.K., as a Post-Doctoral Research Assistant. He was appointed as a Lecturer with Queen's University Belfast in 2003 and was promoted to Senior Lecturer in 2011. He is currently a Professor of cyber security and the Head of the Cyber Security Research Group, University of Derby, U.K. His research interests include cyber security, multimedia security, image and video processing applications, biometrics, and hardware architectures for image and video applications.