Deep Learning for Human Recognition in Motion

1 Introduction

Human recognition technology has unlimited potential uses. This can be in the form of recognizing humans among other objects in a video, facial recognition, behaviour recognition, or activity recognition. However, to get its full potential, the speed and accuracy of recognition are logical requirements if the aim is to imitate the capabilities of human vision. In addition to describing the typical workflow of human image recognition, this review aims to explore the state-of-the-art deep-learning tools and techniques available for image recognition and their applications, discuss the limitations, and propose the aspects that require further investigation. The targeted audience is computer science professionals.

A survey of the literature published in the last ten years was done to identify the relevant article. The literature search was accomplished using the University of Essex Online library and the ACM digital library. In addition, *elicit.org* was used to find appropriate answers to the questions from the literature. Finally, the references of the reviewed articles were also searched for relevant articles.

The review was organized as follows: describing the human/face recognition process, exploring the applications of human/face recognition in motion, explaining a key difference between still images and videos, discussing challenges of human/face recognition in motion, exploring available deep learning frameworks for the human/face detection in motion and their challenges, and lastly discussing conclusions and future directions.

2 Human / Face recognition in Motion

Image recognition of humans in motion is more challenging than identifying fixed objects, which has witnessed significant improvements throughout the years, due to the variable appearance of different human postures, facial expressions, apparel colour, and texture, the effect of the surrounding environment, the occlusion caused by objects or humans, and the availability of training data (Chowdhury et al., 2016; Nguyen et al., 2016; Ali et al., 2021).

The conventional workflow of human recognition in motion is composed of the following steps: image pre-processing, extracting potential image parts, describing those parts, classifying them as human or not, and final processing. Human description comprises features extracted from images like shape, colour, and texture (Nguyen et al., 2016; Ali et al., 2021).

3 Applications of Human / Face recognition in Motion

There are many applications for human/face recognition nowadays. It can be used personally as a biometric identity for personal devices. Also, it can be utilized for security, police, law purposes, surveillance systems at home and public, pedestrians detection, and banking applications (Chen et al., 2014; Ali et al., 2021).

One of the exciting uses of human recognition is “anomaly detection”. This innovative usage is beneficial for detecting abnormal situations in surveillance systems like fights, crimes, and traffic accidents (Jebur et al., 2022).

Anomaly detection can be retrieved from still images or videos. The latter has the advantage of extracting relevant time information needed to differentiate between normal and abnormal behaviours, i.e., violent human interactions probably have higher speeds than normal ones. Anomaly detection can be based on isolated incidents or clustered incidents like fights. Anomaly detection has many practical, real-life applications related to human/face recognition, such as avoiding accidents for self-driving vehicles, automated surveillance systems at home or in public, and violence detection (Ramzan et al., 2019; Khan et al., 2021; Jebur et al., 2022).

Another usage is human action recognition (HAR) or human behaviour recognition (HBR), which has many potential applications in surveillance systems for sports, care centres, classes, and intelligent cities (Dai et al., 2019; Sharma et al., 2021).

4 The key difference between Still Images and Videos

It may not sound new to say that videos contain much more information than still images. However, it is important to realize that along with the documentation of the sequence of images, videos document temporal (related to time) information too. For example, a specific series of frames can be used to forecast what will happen next (Taha et al., 2014). The time component can be utilized to detect anomalies in auto surveillance systems (Jebur et al., 2022).

5 Challenges of the Human / Face Recognition in Motion

One of the most critical challenges of human/face detection is the variation of the subject posture. This leads to images that are more different from than images of other subjects (Ali et al., 2021; Khan et al., 2021).

Variation in the brightness caused by different subjects, cameras, and lighting conditions is another complex problem to overcome (Ali et al., 2021). Occlusion by humans or objects is another challenge. It is essential for auto surveillance systems where capturing humans in cluttered views is unavoidable (Khan et al., 2021). Also, tracing a moving human through several cameras in a surveillance system is a serious challenge (Khan et al., 2021).

Aging is an interesting challenge since it affects the facial surface and shape. Also, it is seldom possible to have updated images of subjects that reflect their most recent appearance in a particular gallery or application (Ali et al., 2021).

Another challenge for some applications, like smart cities, is the need for fast data transfer rate and massive processing power. 5G networks could solve the data transfer rate problem, but the processing power and suitable algorithms could still be a limitation in certain applications (Sankaranarayanan et al., 2008; Dai et al., 2019).

6 Approaches of Human / Face Recognition in Motion

There are conventional approaches and deep learning models. Traditional approaches are based on “global features”, “local features,” or a combination of both, as described by Ali et al. (2021:15-17). This literature review will focus on the deep learning models since they can overcome the challenges explained in the previous section and represent the state-of-the-art in the image recognition (Ali et al., 2021; Jebur et al., 2022).

6.1 Deep Convolutional Networks

The most common deep-learning model used for human image detection is the convolutional neural network (CNN). One of the most significant advantages of CNNs is that they don’t need hand-crafted features as conventional methods. Also, it can better cope with most of the challenges of human detection discussed before, such as variations in brightness, posture, and occlusion (Khan et al., 2021).

A typical CNN approach for videos is composed of two streams; one for temporal information and the other for spatial information (Sharma et al., 2021). Another approach is to use a three-dimensional convolutional network (3D ConvNet) for the concurrent spatiotemporal information handling (Tran et al., 2015).

On the other hand, longer videos may have to be processed to handle the problem of human action recognition. For that purpose, long-term temporal CNN (LTC-CNN) was proposed and showed satisfactory results (Varol et al., 2018).

6.2 Recurrent Neural Network (RNN)

RNN is a type of neural network that is utilized for temporal information. Hence, many human detection methods developed were based on it, such as Long-Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and Neural Turing Machines (NTM) (Sharma et al., 2021).

LSTM is the most popular one of the RNN-based approaches. One example of its usage is for crowd action recognition, where two stages of LTSM are utilized; the first stage is to recognize activity at the individual level, and the second stage is to recognize the group activity by gathering the information from the individual ones (Ibrahim et al., 2016).

6.3 Deep Hashing

Since Hashing has the advantages of speed and low storage, it was implemented in the face recognition methodologies. One example is the deep hashing implemented by Tang et al. (2018). In their work, image recognition was accomplished by utilizing quantization and classification errors to optimize a CNN designed to produce hash codes for face images. Their work proved the accuracy and the performance of such a method.

6.4 Deep learning for Anomaly Detection

Several deep-learning models have been used for anomaly detection. The first one is supervised learning. In this model, the training data is labelled. This can be accomplished using CNNs and recurrent neural networks (RNNs). This type has the advantage of simplicity and performance but requires accurate labels (Jebur et al., 2022).

Another type is the semi-supervised learning model. In this model, training is accomplished on labelled data regarded as normal or wanted. Any other data deviating from the normal will be considered an anomaly. Generative neural networks (GNNs), GRUs, and LSTMs are examples of semi-supervised learning networks. The problem with this framework is that it is prone to errors caused by extraneous features found in the training data (Chalapathy & Chawla, 2019; Jebur et al., 2022).

The next is the unsupervised learning model, where no data labelling is performed. It depends on the big data input and extensive computational power to classify data as normal or abnormal. Other models used for anomaly detection are the transfer learning model, where the learning from previous experiences is transferred to new ones, and the deep active learning model, which requires frequent expert labelling tasks to enhance the model's accuracy. Finally, the deep reinforcement learning model utilizes rewards to enhance learning, like the human learning (Jebur et al., 2022).

6.5 Hybrid models

The last type is deep hybrid models, where several models are utilized to improve the detection result, like CNNs and LSTM, RNNs and CNNS, 3D CNN and LSTM, etc. (Sharma et al., 2021; Jebur et al., 2022).

7 Challenges of the Deep Learning in Human / Face Recognition in Motion

Deep learning has improved the speed and accuracy of human detection in motion. However, some challenges still affect such an approach, which will be summarized in this section.

The first challenge is the inadequate video quality and position of surveillance cameras, which lead to problems in the detection and recognition (Sharma et al., 2021). This may require changing the setup and the type of equipment in a particular setting before deep recognition technology can be applied.

Another challenge is the small training datasets and the limited availability of variable training data, which hinder the ability to generalize a deep learning model for use in different scenarios (Galea & Farrugia, 2018; Ali et al., 2021; Sharma et al., 2021).

A third challenge is a need for adequate computational power to process deep learning algorithms (Sharma et al., 2021). This is especially important if such processing is expected at distant nodes with limited processing power, such as handheld devices or other intelligent systems (Dai et al., 2019).

Finally, due to the sophisticated nature of the neural network, they may be susceptible to adversarial attacks (Papernot et al., 2016; Ali et al., 2021; Poder, 2021).

8 Conclusions and Future Directions

Clearly, the technology of human detection in motion has witnessed key advances. It is being used in many fields at the personal and public levels, whether for security or entertainment. However, as it applies to many areas, research will continue exploring more opportunities for improvements, challenges, and potential uses.

It makes sense to have more sophisticated neural networks to provide smarter results. Nonetheless, more processing power will be needed that may not be readily available for real-life applications (Dai et al., 2019). So, decreasing the complexity of deep neural networks and increasing their efficiency are needed and expected to remain an active research topic.

Although action recognition technology has improved, there is still room for research in the fields of scene comprehension and labelling, processing input from non-fixed cameras, and managing muddled backgrounds (Sankaranarayanan et al., 2008; Sharma et al., 2021).

The last thing to mention here is that facial recognition technology is impressive and may have multiple positive effects on the individual and societal levels. However, It may have ethical implications, mainly represented in protecting personal data and authorizing its use, that require clear regulations for its development and use (Chochia & Nässi, 2021).

References

Ali, W., Tian, W., Din, S. U., Iradukunda, D. & Khan, A. A. (2021) Classical and modern face recognition approaches: a complete review. *Multimedia tools and applications* 80(3)**:** 4825-4880.

Chalapathy, R. & Chawla, S. (2019) Deep Learning for Anomaly Detection: A Survey. Ithaca: Cornell University Library, arXiv.org.

Chen, X., Wei, P., Ke, W., Ye, Q. & Jiao, J. (2014) 'Pedestrian Detection with Deep Convolutional Neural Network'.*Computer Vision - ACCV 2014 Workshops*. Singapore, Singapore, 1-2 November. Springer International Publishing, 354-365.

Chochia, A. & Nässi, T. (2021) Ethics and Emerging Technologies - Facial Recognition. *IDP: Revista d'Internet, Dret i Política***:** 1-12.

Chowdhury, S. A., Uddin, M. N., Kowsar, M. M. S. & Deb, K. (2016) 'Occlusion handling and human detection based on Histogram of Oriented Gradients for automatic video surveillance'.*International Conference on Innovations in Science, Engineering and Technology (ICISET)*. International Islamic University Chittagong, Bangladesh, 28-29 October. IEEE, 1-4.

Dai, C., Liu, X., Lai, J., Li, P. & Chao, H.-C. (2019) Human Behavior Deep Recognition Architecture for Smart City Applications in the 5G Environment. *IEEE network* 33(5)**:** 206-211.

Galea, C. & Farrugia, R. A. (2018) Matching Software-Generated Sketches to Face Photographs With a Very Deep CNN, Morphed Faces, and Transfer Learning. *IEEE transactions on information forensics and security* 13(6)**:** 1421-1431.

Ibrahim, M. S., Muralidharan, S., Zhiwei, D., Vahdat, A. & Mori, G. (2016) ‘A Hierarchical Deep Temporal Model for Group Activity Recognition’.*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA, 27-30 June. IEEE, 1971-1980.

Jebur, S. A., Hussein, K. A., Hoomod, H. K., Alzubaidi, L. & Santamaría, J. (2022) Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance. *Electronics (Basel)* 12(1)**:** 29.

Khan, M. A., Mittal, M., Goyal, L. M. & Roy, S. (2021) A deep survey on supervised learning based human detection and activity classification methods. *Multimedia tools and applications* 80(18)**:** 27867-27923.

Nguyen, D. T., Li, W. & Ogunbona, P. O. (2016) Human detection from images and videos: A survey. *Pattern Recognition* 51(148-175.

Papernot, N., et al. (2016) ‘The Limitations of Deep Learning in Adversarial Settings’.*2016 IEEE European Symposium on Security and Privacy*. Saarbrücken, GERMANY, 2016. IEEE, 372-387.

Poder, E. (2021) Capacity limitations of visual search in deep convolutional neural networks. *Neural Computation (2022)* 34(11).

Ramzan, M., et al. (2019) A Review on State-of-the-Art Violence Detection Techniques. *IEEE access* 7(107560-107575.

Sankaranarayanan, A. C., Veeraraghavan, A. & Chellappa, R. (2008) ‘Object Detection, Tracking and Recognition for Multiple Smart Cameras’.*Proceedings of the IEEE*. New York: IEEE, 1606-1624.

Sharma, V., Gupta, M., Kumar, A. & Mishra, D. (2021) Video Processing Using Deep Learning Techniques: A Systematic Literature Review. *IEEE access* 9(139489-139507.

Taha, A., Zayed, H. H., Khalifa, M. E. & El-Horbaty, E.-S. M. (2014) Exploring Behavior Analysis in Video Surveillance Applications. *International Journal of Computer Applications* 93(22-32.

Tang, J., Li, Z. & Zhu, X. (2018) Supervised deep hashing for scalable face image retrieval. *Pattern recognition* 75(25-32.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Manohar, P. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. Ithaca: Cornell University Library, arXiv.org.

Varol, G., Laptev, I. & Schmid, C. (2018) Long-Term Temporal Convolutions for Action Recognition. *IEEE transactions on pattern analysis and machine intelligence* 40(6)**:** 1510-1517.