Journal of Quality Measurement and Analysis *JQMA* 18(1) 2022, 49-62
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*

A SURVEY ON VIDEO FACE RECOGNITION USING DEEP LEARNING

(Tinjauan Berkenaan Pengecaman Wajah Video Menggunakan Pembelajaran Mendalam)

MUHAMMAD FIRDAUS MUSTAPHA, NUR MAISARAH MOHAMAD, SITI HASLINI AB HAMID, MOHD AZRY ABDUL MALIK & MOHD RAHIMIE MD NOOR

ABSTRACT

The research on facial recognition consists of Still-Image Face Recognition (SIFR) and Video Face Recognition (VFR), is a common subject being debated among researchers since it does not require any touch like other biometric identification, such as fingerprints and palm prints. Various methods have been proposed and developed to solve the problems of face recognition. Convolutional Neural Network (CNN) is one of the deep learning techniques that is suggested for both SIFR and VFR. However, several issues related to VFR have still not been solved. Hence, the objective of this paper is to review VFR using deep learning that specifically focuses on several steps of VFR. The VFR steps consists of six main stages; input video of the face, face anti-spoofing module, face and landmark detection, preprocessing, facial feature extraction and face output that include identification or verification result. A summary of implementation of deep learning within VFR steps is discussed. Finally, some directions for future research are also discussed.

Keywords: convolutional neural network; deep learning; video face recognition

ABSTRAK

Penyelidikan mengenai pengecaman wajah terdiri daripada Pengecaman Wajah Imej Pegun (PWIP) dan Pengecaman Wajah Video (PWV), adalah subjek yang biasa diperdebatkan di kalangan penyelidik kerana tidak memerlukan sentuhan seperti pengenalan biometrik lain, seperti cap jari dan cetakan tapak tangan. Pelbagai kaedah telah dicadangkan dan dibangunkan untuk menyelesaikan masalah pengecaman wajah. Rangkaian Saraf Konvolusional (RSK) adalah salah satu teknik pembelajaran mendalam yang disarankan untuk PWIP dan PWV. Walau bagaimanapun, beberapa masalah yang berkaitan dengan PWV masih belum dapat diselesaikan. Oleh itu, objektif makalah ini adalah untuk mengkaji PWV menggunakan pembelajaran mendalam yang secara khusus menumpukan kepada beberapa langkah PWV. Langkah-langkah PWV terdiri daripada enam peringkat utama; memasukkan video wajah, modul anti-penipuan wajah, pengesanan muka dan mercu tanda, prapemprosesan, pengekstrakan ciri wajah dan output wajah yang merangkumi hasil pengenalan atau pengesahan. Ringkasan pelaksanaan pembelajaran mendalam dalam langkah-langkah PWV telah dibincangkan. Akhir sekali, beberapa hala tuju untuk penyelidikan masa depan juga dibincangkan.

Kata kunci: rangkaian neural konvolusional; pembelajaran mendalam; pengecaman wajah video

1. Introduction

Regarding the development of technology in this information age, the study of the use of biometric systems has continued to be a hot topic for researchers in the field of science and technology. The aim of this biometric system is to define and classify the physical characteristics of a person whose characteristics are unique and different from other persons. Face recognition is a branch of knowledge in the biometric system, and easy to use compared

to other conventional biometric systems. A quick analogy of supporting face recognition the simpler biometric system is unlocking the phone using the face scanning rather than using a fingerprint or thumbprint to unlock it. The human face had details on personality, age, gender, ethnicity and facial expression, representing emotions and mental states. Human face and facial expression analysis is an interdisciplinary field of study including psychology and neuroscience (Martinez *et al.* 2019). Face recognition can be split into two sections which are still-image face recognition (SIFR) and video face recognition (VFR). Basically, the SIFR systems are attempting to identify an individual by using a physical look. Otherwise, VFR systems use physical changes in over the time or dynamic of the face that discussed in (Taskiran *et al.* 2020). VFR is more difficult compared to SIFR due to a much greater amount of data to be processed and significant inter- or intra-class variations caused by low video quality, motion blur, occlutions and frequent changes in scene (Zheng *et al.* 2020).

In addition, face recognition has been widely applied to everyday lives with the advancement of imaging technology and computer hardware. In recent years, the demands for face recognition are growing rapidly as reported in (Guo & Zhang 2019). While these related surveys highlighted methods of handling pose (Manju & Radha 2020), illumination, expression (Hassouneh *et al.* 2020), occlusion (Lahasan *et al.* 2019), infrared (Guo *et al.* 2017), single-modal and multi-modal (Imani & Montazer 2019), video, 3D, heterogeneous face matching, etc., mostly concentrated on conventional methods, and several of them applied deep learning methods.

Deep learning is a form of machine learning (ML), a main branch of the Artificial Intelligence (AI) domain, that concerned with the science and engineering of machines with features of human intelligence (Meijering 2020). There are various deep learning techniques in ML such as Convolutional Neural Network (CNN), Deep Reinforcement Learning (DRL), Recurrent Neural Network (RNN), Deep Neural Network (DNN), and Deep Belief Network (DBN) (Jauro *et al.* 2020). CNN is the most popular techniques and has been applied for VFR and SIFR (Guo & Zhang 2019).

Therefore, the objective of this paper is to review VFR using deep learning that specifically focuses on several steps of VFR. The following section is to be arranged as follows. Section 2 explains the related work on face recognition and deep learning. Section 3 describes the preliminary data analysis of the paper. Section 4 discusses the analysis on VFR steps. Section 5 contains the explanation of the application of feature extraction in CNN architecture of face recognition. Finally, Section 6 gives the conclusion that consist of discussion and future works on VFR using deep learning.

2. Literature Review

Since the 9/11 incidents in the United States and with the current security threats, surveillance systems focused on face recognition are gaining growing interest. Real-time detection can be used for example to surveillance control that discussed in (Luo *et al.* 2020). Real-time detection consists of VFR and SIFR. VFR has the key advantage, as compared with more conventional SIFR, of using numerous cases of almost the same entity in sequential frames for identification. In the case of SIFR, the device has only one input image for making a decision if the individual is in the database or not. If the image is not appropriate for recognition (due to face orientation, voice, clarity or facial occlusions), it would most likely be incorrect to recognize the image. On the other hand, there are a few frames in the video image that can be examined to provide better accuracy in the recognition (Jauro *et al.* 2020). Although certain frames are not appropriate for recognition there is a strong possibility that some of them will perform and a strong level of confidence could be in the decision taken

(Corcoran 2011). If a face is identified, tracking strategies remain recognizable throughout the scene.

The VFR has arisen as an exceedingly relevant field of study. Even worse, the bulk of current literature on face recognition centers on matching the SIFR, and research on VFR is still in its inception. The main problem is to create an adequate visual representation of video faces so that informations can be easily shared across multiple frames. VFR is substantially more challenging than SIFR (Cheng *et al.* 2018). VFR offers several types of data for face recognition, sampling and modeling, video frame image quality seems to be considerably lower, and faces reveal much richer variations, as video acquisition may be let alone restrictive. Topics in videos, for example, are typically mobile, resulting in significant out-of-focus blur, mobile, motion blur, and a wide variety of pose variations. In addition, mobile cameras and surveillance frequently become low-cost (and thus low-quality) devices which further exacerbate video frame issues (Guo & Zhang 2019). Current developments in face recognition have appeared to neglect the peculiarities of images when applying SIFR to VFR techniques (Li & Hua 2015; Schroff *et al.* 2015; Sun *et al.* 2015). A big problem in VFR, such as extreme image blur, remains still unresolved (Beveridge *et al.* 2015). An important reason for this, is that there is still a lack of large amounts of real-world video training data and current still image sources are typically blur-free. On the other hand, while occlusion and pose variations are partially solved by ensemble modelling in SIFR (Liu *et al.* 2015), the technique may not be applied explicitly to VFR. Nevertheless, a large quantity of footage is continually being collected due to the increasing number of CCTV cameras installed and the convenient availability of video recordings. VFR typically provides more details, e.g., temporal and multi-view information, relative to SIFR. The prevalence of videos has far-reaching implications for society in terms of defense and law enforcement. Construction of security devices coupled with face recognition techniques is particularly attractive in order to instantly recognize subjects of interest.

VFR lends itself well to the application of deep learning because deep learning has strength and accuracy to handle vast quantities of data, as well as gained interest in a variety of fields (Wani *et al.* 2020). As a consequence, in recent years, deep learning algorithms have been applied to solve numerous issues in emerging cloud computing architectures including anomaly detection, cyber security, object detection, street cleanliness, food recognition, smart vehicles, re-identification of people, fruit classification, etc. (Li *et al.* 2017*b*; Parchami *et al.* 2017; Pranav & Manikandan 2020). Several deep learning-based solutions have recently been demonstrated for implementing new face representations instantly from training data through CNN and non-linear feature mappings (Jauro *et al.* 2020).

Deep learning is a main branch of the Artificial Intelligence domain that concerned with the science and engineering of machines with features of human intelligence (Meijering 2020). There are many deep learning techniques can be applied for VFR such as CNN, DRL, RNN, and DBN (Jauro *et al.* 2020). CNN is one of the deep learning techniques that is gaining popularity among researchers and this technique was established for the analysis of visual data, such as images and videos.

CNN has been applied by many researchers to solve various problems in face recognition (Li & Hua 2015; Parchami *et al.* 2017; Pranav & Manikandan 2020). For example, Pranav and Manikandan (2020) have designed and evaluated real-time face recognition using CNN to easily adapted for various consumer applications such as attendance system and device control. Their proposed system achieved higher accuracy level for standard dataset and real-time input. Works by Silva and Jung (2020) revealed the application of CNN in an end-to-end Automatic License Plate Recognition method. It was tested with publicly available datasets containing Brazilian and European license plates and achieved accuracy rates better than competitive academic methods and a commercial system. Similarly with HyperFace model

that was proposed by Ranjan *et al.* (2019) able to capture both global and local information in faces and performs significantly better than many competitive algorithms for each of these four tasks which are face detection, pose estimation, gender recognition and landmark localization using CNN.

CNN has its advantages compared to other techniques because many vision tasks have benefited from the robust, discriminative representation learned via CNN and the performance has been enhanced significantly (Guo & Zhang 2019). In this regard, this paper examines the steps in VFR using deep learning that specifically uses CNN techniques.

3. Bibliometric Data Analysis

The documents related to VFR using deep learning algorithms is derived for the interval of past few years and involve collecting extensive literature on Dec, 2020. The data since 2010 to the year 2020 is depicted in Figure 1. According to the Figure 1, the document that selected includes about 55 papers, and most of the papers within the recent six years. The keywords used for this review are "face recognition", "video face recognition", "deep learning", and "convolutional neural network". Moreover, the academic databases used to extract the documents is presented in Table 1. Based on the table, there are five different sources of document. Most of the documents are gathered from IEEE source with 23 documents, followed by ScienceDirect, Scopus, Google Scholar and book.

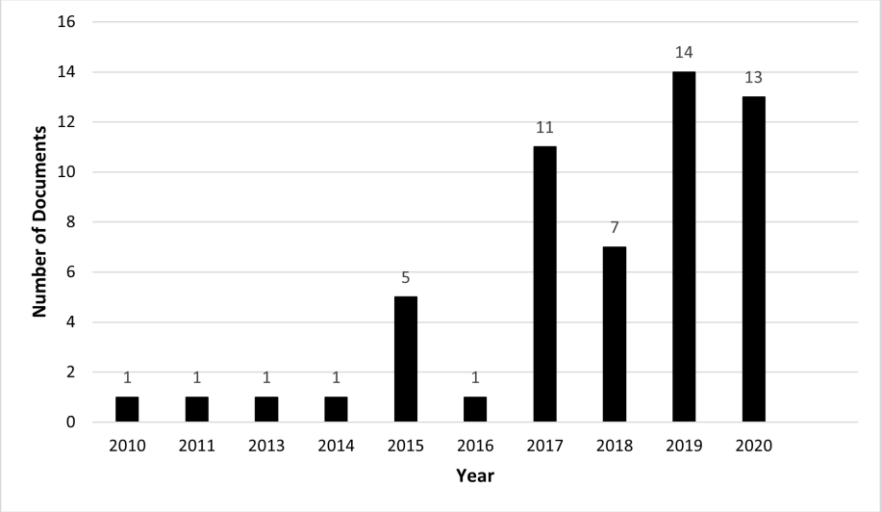

Figure 1: Document by year

Table 1: Document by source

Source	Number of documents
Book	1
Google Scholar	6
IEEE	23
ScienceDirect	15
Scopus	10

Table 2 shows the top five countries that published the selected documents into this paper. Most of the documents are published by China (14), followed by United States (8), Turkey

(4), India (4) and United Kingdom (3). The rest of the countries are Korea, Czechia, Ireland, Kuwait, Iran, Nigeria, France, Italy, Malaysia, Australia, Thailand, Brazil, and Singapore. These countries have published at least one article.

Table 2: Top 5 countries publishing the selected articles

Country	Number of documents
China	14
United States	8
Turkey	4
India	4
United Kingdom	3

.

4. Analysis on Video Face Recognition Steps

The development of VFR using deep learning technique consists of six main steps; facial input video, face anti-spoofing module, face and landmark detection, preprocessing, facial feature extraction, and facial identification or verification output as shown in Figure 2 (Taskiran *et al*. 2020).

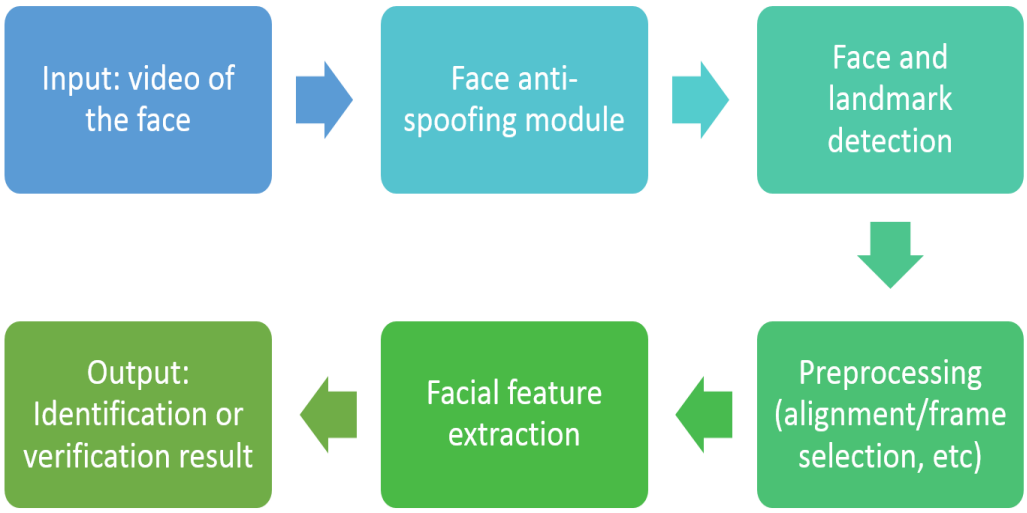

Figure 2: Main step of VFR

4.1. *Input video of the face*

VFR has attracted attention from many researchers and therefore many video face datasets were released. Table 3 illustrates a list of datasets for video faces and most of those are open to the public. Youtube Face (YTF) and Point and Shoot Challenge (PaSC) have been used to examine the performance of numerous deep models on recognition (Guo & Zhang 2019). Although iQIYI-VID contains the largest video dataset, but it is a newly released dataset and not many researchers applied it to examine the performance of deep models on VFR.

Table 3: Example of video face datasets used for VFR

Dataset	Videos	Identities	Description
ACVF-2014	201	133	Multiple subjects in frame; use handheld cameras
Celebrities-1000	7021	1000	Covering illuminations, poses, etc.
ChokePoint	48	29	Video surveillance dataset; 64,204 still images
CSCRV	193	160	Video; with open-set protocol
ESOGU-286	2280	286	764k frames; set-based FR
Faces96	152	152	Significant head variations
FIA	6470	180	Captured by 6 synchronized cameras from 3 different angles
Honda	59	20	Large pose/expression variations; 400 frame/video
iQIYI-VID	600k	5000	From 400k hours of online videos
McGillFaces	60	60	Real-world face Video
PaSC	2802	293	Still + Video; collected at different locations, poses, and distances
SN-Flip	28	190	Multiple subjects in frame; less motion
UMDFaces-Videos	22075	3107	Video; from YouTube
YTC	1910	47	High compression rate; large variations; from YouTube
YTF	3425	1595	Low resolution, motion blur; from YouTube

4.2. *Face anti-spoofing module*

Face-recognition anti-spoofing typically involves the monitoring of liveness or appearance of an attack, can directly be achieved by sensing facial gestures such as eye blinking (Ali *et al.* 2018), changes of facial expression, movement of the mouth, or movement of the head (Lagorio *et al.* 2013). Ability to detect the heart rate from a face video is another way to detect liveness (Wang *et al.* 2019). This technique is known as remote photoplethysmography (non-contact), which uses the slight color patterns of the skin that happen each time the heart beats and pumps blood to the body (Demirezen & Erdem 2018). Certain counter-measures may include different biometric types, for instance voice and gait. After all, multi-modal systems are designed quite hard to spoof than uni-modal systems (Killioğlu *et al.* 2017). Further details on countering 2D photo spoofing threats that could be discovered in (Chrzan 2014).

CNN is one of the deep learning techniques that has recently become an important method for anti-spoofing (Shao *et al.* 2019). In (Nagpal & Dubey 2019), the results of various CNN architectures for face anti-spoofing is analysed. In (Liu *et al.* 2019), the deep tree learning (DTL) method was introduced for the zero-shot face anti-spoofing (ZSFA) technique. ZSFA is the detection of spoof attacks that do not occur in training data, such as transparent mask attacks or incomplete paper attacks. Moreover, a spatio-temporal anti spoofing network (STASN) was proposed in (Yang *et al.* 2019), which can rely on subtle cues, including border and moire patterns, to identify spoof face. Then, data collection and synthesis methods were also provided. A multi-modal anti-spoofing test was also newly conducted using the CASIA-SURF multi-modal dataset in (Zhang *et al.* 2019). This study outlines the findings of the most active teams and presented valuable insights on possible avenues for new researcher.

Face anti-spoofing techniques are used to detect spoof attacks before performing recognition in order to use face recognition securely. Spoof attacks can be carried out through a multitude of channels. Replaying videos or images on computer screens, also known as replay attack, and printing a photograph, generally known as print attack, are the most typical of spoof attack.

4.3. *Face and landmark detection*

Face detection is the measurement of the face boundary in a video frame (Taskiran *et al.* 2020). If the frames contain numerous images, they will all be detected. The face detection must be robust to variations in posture, lighting and scale and should remove the context as far as possible. Lately, deep-learning face detectors have reported positive results (Ranjan *et al.* 2018). Inside a recent method, faster Region-based CNN (R-CNN), that uses the regional proposal technique, was originally introduced for object detection (Wu *et al.* 2019). There are a few other deep-learning face detection methods that use the sliding-window concept (Peng & Gopalakrishnan 2019). Commonly utilized successfully for face detection was the single shot detector (SSD), which was originally introduced for face detection (Yang *et al.* 2017). Once the face is detected, facial landmarks on the face (eyebrows, eye corners, mouth, nose tip, etc.) that can also be used for face alignment is expected as depicted in Figure 3. Multi-task learning methods were developed, with additional tasks containing prediction and gender recognition (Ranjan *et al.* 2019).

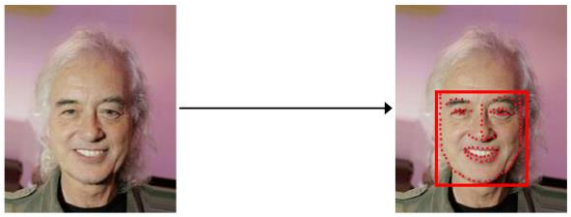

Figure 3: Face and landmark detection step

4.4. *Preprocessing*

Preprocessing is done on an image or video and may require contrast enhancement, alignment, video frame selection, noise reduction, or related activities (Taskiran *et al.* 2020). Deep learning based face recognition approaches present a specific robustness in identifying facial images in unregulated environments with different facial expressions, lighting, and posture. Fortunately, a previous study (Ghazi & Ekenel 2016) observed that various facial expressions, lighting, and exposure had adverse effects on network performance and indicated the need for face pre-processing to improve performance. The face pre-processing methods are categorized as "one-to-many augmentation" and "many-to-one normalization" (Taskiran *et al.* 2020), as shown in Table 4.

Table 4: Types of face pre-processing methods

Face Pre-processing	Goals	Methods
One to many	Create images in various poses from a single image during training to make the deep CNN pose-invariant.	Data augmentation (Lv *et al.* 2017), 3D model (Zhou & Xiao 2018), 2D deep model (Soltana *et al.* 2010), Autoencoders (Li *et al.* 2017*a*)
Many to one	Create the canonical view of a face image by combining multiple face images taken from various angles in an uncontrolled environment.	CNN (Pitaloka *et al.* 2017), GAN (Cho *et al.* 2018)

In one-to-many augmentation pre-processing approach, the aim is to make the deep CNN pose-invariant by generating images indifferent poses from a single image during training. This is done since it is expensive and time-consuming to collect large numbers of images for creating a training database.

In many-to-one normalization pre-processing approach, the goal is to recover the canonical view of face images from one or many images of a nonfrontal view; then, VFR can be performed as if it were under controlled conditions.

Note that this paper mainly focus on face processing method designed for deep learning like CNN, since CNN is widely used by researchers to overcome the challenges in VFR applications and other variations can be solved by the similar methods.

4.5. *Facial feature extraction*

Two main methods that can be used to extract facial features from video are set-based and sequence-based (Taskiran *et al.* 2020) as depicted in Figure 4. This paper focuses on image set-based approach to face recognition which video frames are viewed as a collection of image samples and the temporal order is not addressed. Set-based approach may be defined as a method that use fusion before and after matching. Fusion before matching involves join in the features acquired from each face image before the process of recognition. Fusion after matching technique blends the outcomes of the recognition of each image. This confederation can be achieved by a rank, score, or a decision level merger.

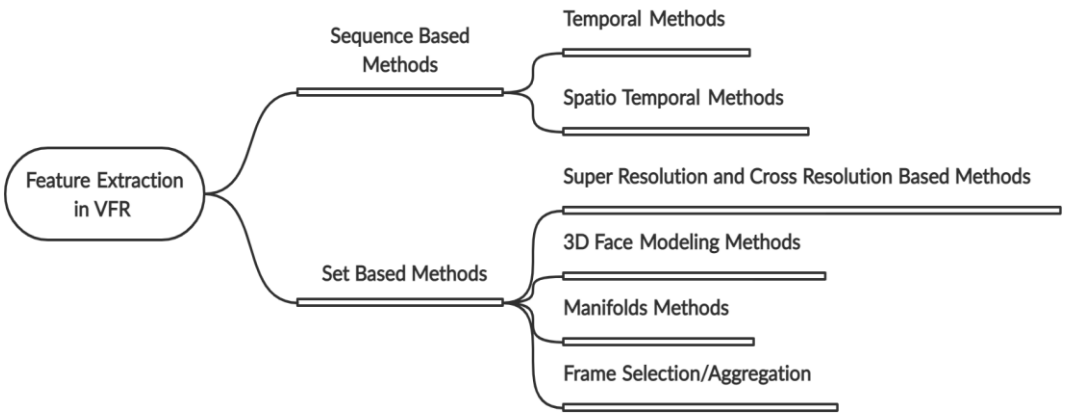

Figure 4: Feature extraction in VFR

Super-resolution and cross-resolution have been recommended for VFR. Super-resolution is the process by which low-resolution images acquired from surveillance cameras are used to gather information found in a high-resolution image. Preferably from just increasing the resolution and extracting a feature from a high-resolution image, the super-resolution method was utilized to the feature vectors provided from several low-resolution images, and only facial information was collected (Taskiran *et al.* 2020). The existing super resolution methods use several kinds of interpolation techniques. Nowadays, CNN can be used to extract facial features. For example, the super-resolution convolutional neural network (SRCNN) (Ahn *et al.* 2019) has been introduced. Recently, a common approach to cross-resolution face recognition has been used in (Ahn *et al.* 2019; Massoli *et al.* 2020). Initially, discriminatory

features that are robust to pose variations are learned in low-resolution and high-resolution spaces by multilayer locality-restricted structural orthogonal regression of Procrustes. Subsequently, the recognition is done using other solution features. Cross-resolution face recognition handles the problem of matching face images with distinct resolution. State-of-the-art CNN based method has mentioned convincing performances on standard face recognition issues (Fu *et al.* 2017).

4.6. *Identification or verification*

The output of face recognition can be done using an identification or verification (authentication) approach as summarized in Table 5. Face identification is a one-to-many mapping for a given face against a database of known faces or identities. On the other hand, face verification is a one-to-one mapping of a given face against a known identity in the database. Identity authentication elevates verification to a higher level, which is especially important when dealing with digital purchasing. Deep learning approaches have very high recognition accuracy ratings across a large amount of high-quality images captured in unregulated environments for both face identification and verification, but their robustness under adverse circumstances is being studied (Grm *et al.* 2018). Nevertheless, lower recognition accuracy has been reported where there are severe illumination differences, low resolution images or noise (Grm *et al.* 2018). Hence, under adverse circumstances, VFR approaches may offer beneficial facial dynamics information. The main aim is to enhance the image by applying super resolution algorithms or cross-resolution algorithms to low resolution face images in order to improve the efficiency of face recognition systems. Since there is a large amount of data collected from surveillance cameras, low-resolution image recognition is a significant and difficult research topic to investigate (Masi *et al.* 2017).

Table 5: Summary of identification and verification approach

	Identification	Verification
Known as	1:N matching problem	1:1 matching problem
Explanation	The unknown face is compared with all the faces in the database of known identities	The identity of the query face is compared with the face data of the claimed identity in the databases
Type of task / result	• closed-set - person is known to be in the database • open-set – person is unknown to be in the database	• confirmed • rejected

5. Convolutional Neural Network for Face Recognition

CNN is a deep learning algorithm that is commonly used by researchers (Kortli *et al.* 2020). CNN can classify, analyse, and process high-dimensional patterns with the right network architecture, making it an extremely valuable tool in computer vision (Arachchilage 2020). Recent advancements in CNN architectures in recognising human faces have resulted in excellent performance of several CNN-based models in recognising faces in video. To learn spatial hierarchies of image features, this network typically employs an activation function and training algorithms (Zheng *et al.* 2020). As a result, images are used as input labels, and training is carried out automatically. Figure 5 shows the uses of feature extraction in CNN architecture of face recognition.

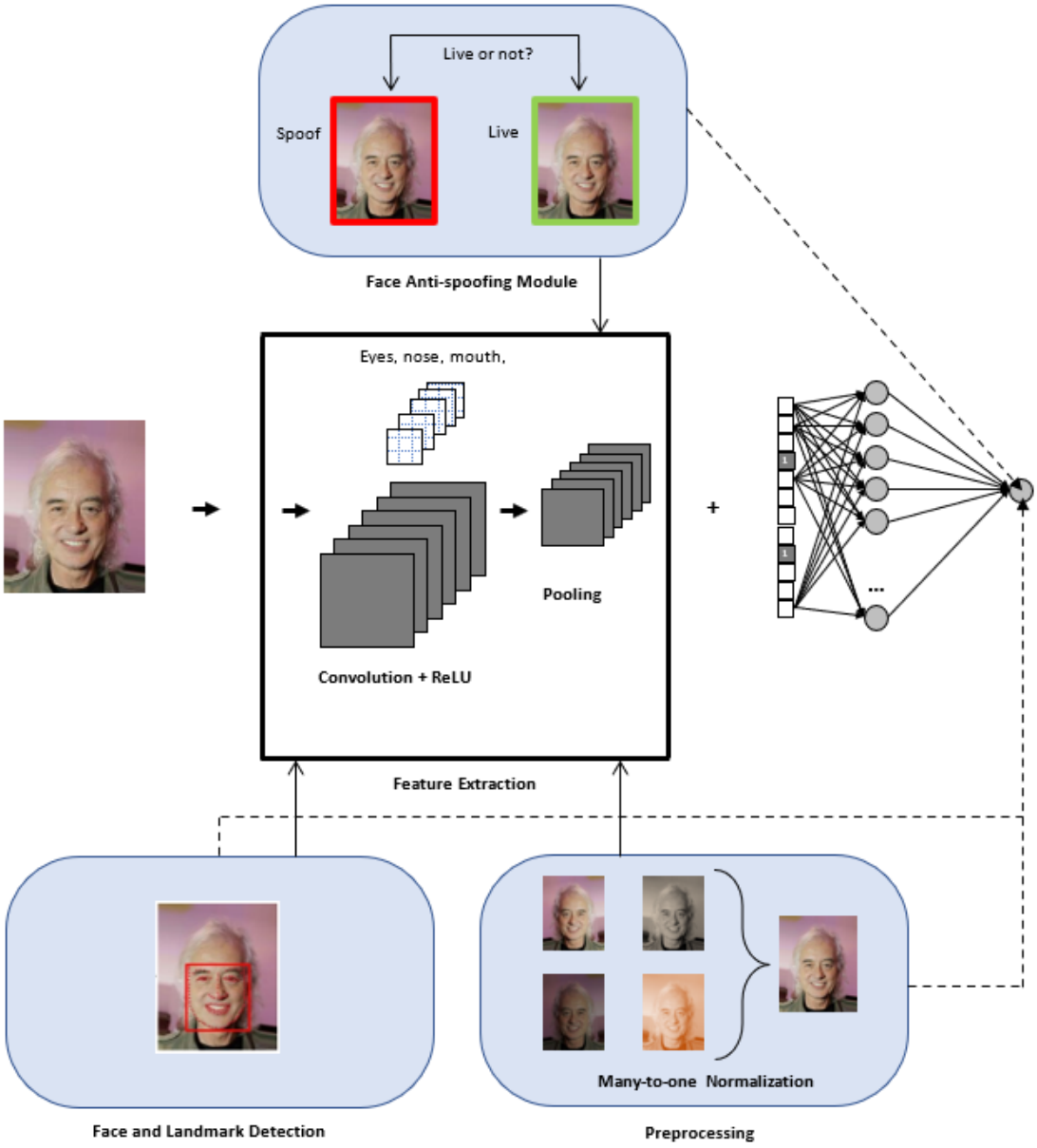

Figure 5: Example of CNN architecture in face recognition

The stage of Face Anti-Spoofing Module, Feature Extraction, Face and Landmark Detection, and Preprocessing have been discussed in the previous section. The main idea behind the CNN is feature extraction. Figure 5 shows the uses of feature extraction in CNN architecture to detect face anti-spoofing, face and landmark detection, preprocessing which is many-to-one normalization and for image classification. The feature extraction in CNN uses convolutional and pooling layers along with input and output layers. At this stage, basically the feature extraction has several layers of Convolution and Pooling depending on the purpose of the experiment being conducted (Taskiran *et al.* 2020). The first Convolution layer, will start with a small part of the face that will build up the eyes, nose, mouth, etc. When looking to solve problems in deep learning, it is non-linear in nature (Zheng *et al.* 2020). Therefore,

the convolutional layer is usually accompanied by the Rectified Linear Unit (ReLU) layer (Ahn *et al.* 2019; Massoli *et al.* 2020). It also speeds up training and is faster to count. The advantage of the Pooling layer is it reduces dimensions and calculations (Ahn *et al.* 2019; Massoli *et al.* 2020). It reduces redundant installation and makes the model tolerant of every minor distortion. Pooling layers usually have two types, namely Max Pooling and Average Pooling. The number of layers depends on the architecture, the data, and the performance required from the model. CNN are not prone to overfitting due to a reduction in weights and the number of neurons caused by the convolutional layer and pooling layer, respectively (Jauro *et al.* 2020). After that, the second stage is a fully-connected layers that are used for classification (image classification) and identification or verification for (face recognition) which has been elaborated in section 4.6.

6. Conclusion and Future Work

This paper has presented a survey on video face recognition using deep learning. This study has discussed six main stages of VFR steps and also CNN for face recognition. The finding of this study will redound to the benefit of society considering that video face recognition system plays an important role in face recognition and computer vision. The ultimate goal is to create an advanced face recognition system that can exceed the human vision system. The stages of VFR steps using deep learning technique were summarized as follows;

(1) Input video of the face can be gathered manually or get from public video faces dataset
(2) Face anti-spoofing module ensures the security of the system by employing presentation or adversarial attack detection
(3) Face and facial landmarks are detected in each of video frame
(4) Preprocessing is performed on the video which may consist of video frame selection, noise reduction, contrast enhancement, alignment or similar operations
(5) Facial feature extraction, which used set-based approaches super-resolution and cross resolution provides promising performance on low-resolution of the face in each video frame
(6) Face identification or verification can be performed as the output of VFR

In conclusion, deep learning technique can be applied in most of the steps in VFR such as face anti-spoofing module, face and landmark detection and also facial feature extraction. The implementation of deep learning within VFR steps can trigger a significant impact to accuracy and processing speed (Franc & Čech 2018; Pranav & Manikandan 2020). Deep learning technique in VFR is capable of utilising very large datasets of faces and of training rich and compact representations of faces, allowing computational models to succeed first as well as later to outperform human face recognition capabilities. However, there are some challenges for VFR under adverse circumstances such as low resolution images. Therefore, future research work on VFR using deep learning will be investigated for a prolonged time and more research will be continuosly proposed in the literature to improve the accuracy performance especially under adverse circumstance. Several improvements should be hightlighted in the VFR steps such as face anti-spoofing module, preprocessing, face and landmark detection as well as facial feature extraction.

Acknowledgments

This work was supported by Fundamental Research Grant Scheme (FRGS) under Ministry of Education (MOE) with grant number 600-IRMI/FRGS 5/3 (234/2019).

References

Ahn H., Chung B. & Yim C. 2019. Super-resolution convolutional neural networks using modified and bilateral ReLU. *ICEIC 2019 - International Conference on Electronics, Information, and Communication*, pp. 30–33.

Ali A., Hoque S. & Deravi F. 2018. Gaze stability for liveness detection. *Pattern Analysis and Applications* **21**(2): 437–449.

Arachchilage S.P.K.W. 2020. Deep-learned faces : a survey. *EURASIP Journal on Image and Video Processing* **2020**: 1–33.

Beveridge J.R., Zhang H., Draper B.A., Flynn P.J., Feng Z., Huber P., Kittler J., Huang Z., Li S., Li Y., Kan M., Wang R., Shan S., Chen X., Li H., Hua G., Struc V., Krizaj J., Ding C., Tao D. & Phillips P.J. 2015. Report on the FG 2015 Video Person Recognition Evaluation. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2015).*

Cheng W.C., Wu T.Y. & Li D.W. 2018. Ensemble convolutional neural networks for face recognition. *ACM International Conference Proceeding Series* **40**(4): 1002–1014.

Cho S., Kim D., Yoo S. & Sohn C.B. 2018. Generative Adversarial Network-Based Face Recognition Dataset Generation. *International Journal of Applied Engineering Research* **13**(22): 15734–15739.

Chrzan B.M. 2014. Liveness detection for face recognition. Master Thesis. Masaryk University.

Corcoran P. (ed.) 2011. *New Approaches to Characterization and Recognition of Faces*. Rijeka: IntechOpen.

Demirezen H. & Erdem C.E. 2018. Remote photoplethysmography using nonlinear mode decomposition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 1060–1064. IEEE.

Franc V. & Čech J. 2018. Learning CNNs from weakly annotated facial images. *Image and Vision Computing* **77**: 10–20.

Fu T.C., Chiu W.C. & Wang Y.C.F. 2017. Learning guided convolutional neural networks for cross-resolution face recognition. *IEEE International Workshop on Machine Learning for Signal Processing, MLSP* pp. 1–6.

Ghazi M.M. & Ekenel H.K. 2016. A comprehensive analysis of deep learning based representation for face recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 102–109.

Grm K., Struc V., Artiges A., Caron M. & Ekenel H.K. 2018. Strengths and weaknesses of deep learning models for face recognition against image degradations. *IET Biometrics* **7**(1): 81–89.

Guo G. & Zhang N. 2019. A survey on deep learning based face recognition. *Computer Vision and Image Understanding* **189**(July): 102805.

Guo K., Wu S. & Xu Y. 2017. Face recognition using both visible light image and near-infrared image and a deep network. *CAAI Transactions on Intelligence Technology* **2**(1): 39–47.

Hassouneh A., Mutawa A.M. & Murugappan M. 2020. Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods. *Informatics in Medicine Unlocked* **20**: 100372.

Imani M. & Montazer G.A. 2019. A survey of emotion recognition methods with emphasis on E-Learning environments. *Journal of Network and Computer Applications* **147**: 102423.

Jauro F., Chiroma H., Gital A.Y., Almutairi M., Abdulhamid S.M. & Abawajy J.H. 2020. Deep learning architectures in emerging cloud computing architectures: Recent development, challenges and next research trend. *Applied Soft Computing Journal* **96**: 106582.

Killioğlu M., Taşkiran M. & Kahraman N. 2017. Anti-spoofing in face recognition with liveness detection using pupil tracking. *SAMI 2017 - IEEE 15th International Symposium on Applied Machine Intelligence and Informatics, Proceedings*, pp. 87–92.

Kortli Y., Jridi M., Al Falou A. & Atri M. 2020. Face recognition systems: A survey. *Sensors* **20**(2): 342.

Lagorio A., Tistarelli M., Cadoni M., Fookes C. & Sridharan S. 2013. Liveness detection based on 3D face shape analysis. *2013 International Workshop on Biometrics and Forensics, IWBF 2013*, pp. 6–9.

Lahasan B., Lutfi S.L. & San-Segundo R. 2019. A survey on techniques to handle face recognition challenges: occlusion, single sample per subject and expression. *Artificial Intelligence Review* **52**(2): 949–979.

Li F., Gao X. & Wang L. 2017*a*. Face recognition based on deep autoencoder networks with dropout. *2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2017)*, pp. 243–246.

Li H. & Hua G. 2015. Hierarchical-PEP model for real-world face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4055–4064.

Li J., Zhang D., Zhang J., Zhang J., Li T., Xia Y., Yan Q. & Xun L. 2017*b*. Facial expression recognition with faster R-CNN. *Procedia Computer Science* **107**: 135–140.

Liu J., Deng Y., Bai T., Wei Z. & Huang C. 2015. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv*: 1506.07310.

Liu Y., Stehouwer J., Jourabloo A. & Liu X. 2019. Deep tree learning for zero-shot face anti-spoofing. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4675–4684.

Luo H., Liu J., Fang W., Love P.E.D., Yu Q. & Lu Z. 2020. Real-time smart video surveillance to manage safety: A case study of a transport mega-project. *Advanced Engineering Informatics* **45**: 101100.

Lv J.J., Shao X.H., Huang J.S., Zhou X.D. & Zhou X. 2017. Data augmentation for face recognition. *Neurocomputing* **230**: 184–196.

Manju D. & Radha V. 2020. A novel approach for pose invariant face recognition in surveillance videos. *Procedia Computer Science* **167**: 890–899.

Martinez B., Valstar M.F., Jiang B. & Pantic M. 2019. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing* **10**(3): 325–347.

Masi I., Hassner T., Tran A.T. & Medioni G. 2017. Rapid synthesis of massive face sets for improved face recognition. *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, pp. 604–611.

Massoli F.V., Amato G. & Falchi F. 2020. Cross-resolution learning for face recognition. *Image and Vision Computing* **99**: 103927.

Meijering E. 2020. A bird's-eye view of deep learning in bioimage analysis. *Computational and Structural Biotechnology Journal* **18**: 2312–2325.

Nagpal C. & Dubey S.R. 2019. A performance evaluation of convolutional neural networks for face anti spoofing. *Proceedings of the International Joint Conference on Neural Networks*, pp. 1–8.

Parchami M., Bashbaghi S. & Granger E. 2017. Video-based face recognition using ensemble of Haar-like deep convolutional neural networks. *Proceedings of the International Joint Conference on Neural Networks*, pp. 4625–4632.

Peng B. & Gopalakrishnan A.K. 2019. A face detection framework based on deep cascaded full convolutional neural networks. *2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019*, pp. 47–51.

Pitaloka D.A., Wulandari A., Basaruddin T. & Liliana D.Y. 2017. Enhancing CNN with preprocessing stage in automatic emotion recognition. *Procedia Computer Science* **116**: 523–529.

Pranav K.B. & Manikandan J. 2020. Design and evaluation of a real-time face recognition system using convolutional neural networks. *Procedia Computer Science* **171**: 1651–1659.

Ranjan R., Sankaranarayanan S., Bansal A., Bodla N., Chen J.C., Patel V.M., Castillo C.D. & Chellappa R. 2018. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine* **35**(1): 66–83.

Ranjan R., Patel V.M. & Chellappa R. 2019. HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(1): 121–135.

Schroff F., Kalenichenko D. & Philbin J. 2015. FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 815–823.

Shao R., Lan X. & Yuen P.C. 2019. Joint discriminative learning of deep dynamic textures for 3D mask face anti-spoofing. *IEEE Transactions on Information Forensics and Security* **14**(4): 923–938.

Silva S.M. & Jung C.R. 2020. Real-time license plate detection and recognition using deep convolutional neural networks. *Journal of Visual Communication and Image Representation* **71**: 102773.

Soltana W.B., Huang D., Ardabilian M., Chen L. & Amar C.B. 2010. Comparison of 2D/3D features and their adaptive score level fusion for 3D face recognition. *International Symposium on 3D Data Processing, Visualization and Transmission* in Paris, France.

Sun Y., Wang X. & Tang X. 2015. Deeply learned face representations are sparse, selective, and robust. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2892–2900.

Taskiran M., Kahraman N. & Erdem C.E. 2020. Face recognition: Past, present and future (a review). *Digital Signal Processing: A Review Journal* **106**: 102809.

Wang J., Chen Y., Hao S., Peng X. & Hu L. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* **119**: 3–11.

Wani M.A., Bhat F.A., Afzal S. & Khan A.I. 2020. *Advances in Deep Learning*. Singapore: Springer.

Wu W., Yin Y., Wang X. & Xu D. 2019. Face detection with different scales based on faster R-CNN. *IEEE Transactions on Cybernetics* **49**(11): 4017–4028.

Yang S., Xiong Y., Loy C.C. & Tang X. 2017. Face detection through scale-friendly deep convolutional networks. *arXiv:1706.02863*

Yang X., Luo W., Bao L., Gao Y., Gong D., Zheng S., Li Z. & Liu W. 2019. Face anti-spoofing: Model matters, so does data. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3502–3511.

Zhang S., Wang X., Liu A., Zhao C., Wan J., Escalera S., Shi H., Wang Z. & Li S. Z. 2019. A dataset and benchmark for large-scale multi-modal face anti-spoofing. *Proceedings of the IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition, pp. 919–928.
Zheng J., Ranjan R., Chen C.H., Chen J.C., Castillo C.D. & Chellappa R. 2020. An automatic system for unconstrained video-based face recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science* **2**(3): 194–209.
Zhou S. & Xiao S. 2018. 3D face recognition: a survey. *Human-centric Computing and Information Sciences* **8**(1).

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Bukit Ilmu, 18500 Machang
Kelantan, Malaysia
E-mail: mdfirdaus@uitm.edu.my[*]*, mohdr697@uitm.edu.my, azry056@uitm.edu.my*

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
Jalan Ilmu 1/1, 40450 Shah Alam
Selangor, Malaysia
E-mail: maisarah97.mohamad@gmail.com

FH Training Center
Lot 21, Kampung Alor Pasir
16800 Pasir Puteh
Kelantan, Malaysia
E-mail: cthaslini@gmail.com

Received: 1 December 2021
Accepted: 16 January 2022

[*]Corresponding author