Received: 18 December 2018 | Revised: 24 June 2020 | Accepted: 30 July 2020

DOI: 10.1049/tje2.12001

The Journal of Engineering

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

A benchmark of dynamic versus static methods for facial action unit detection

L. Alharbawee[1,2] | **N. Pugeault**[1,3]

[1] College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

[2] College of Computer Sciences and Mathematics, Statistics and Informatics department, University of Mosul, Mosul, Iraq

[3] School of Computing Science, University of Glasgow, Glasgow, Scotland

Correspondence
L. Alharbawee, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK.
Email: la315@exeter.ac.uk

Funding information
University of Glasgow

Abstract

Action Units activation is a set of local individual facial muscle parts that occur in time constituting a natural facial expression event. AUs occurrence activation detection can be inferred as temporally consecutive evolving movements of these parts. Detecting AUs automatically can provide explicit benefits since it considers both static and dynamic facial features. Our work is divided into three contributions: first, we extracted the features from Local Binary Patterns, Local Phase Quantisation, and dynamic texture descriptor LPQ-TOP with two distinct leveraged network models from different CNN architectures for local deep visual learning for AU image analysis. Second, cascading the LPQTOP feature vector with Long Short-Term Memory is used for coding longer term temporal information. Next, we discovered the importance of stacking LSTM on top of CNN for learning temporal information in combining the spatially and temporally schemes simultaneously. Also, we hypothesised that using an unsupervised Slow Feature Analysis method is able to leach invariant information from dynamic textures. Third, we compared continuous scoring predictions between LPQTOP and SVM, LPQTOP with LSTM, and AlexNet. A competitive substantial performance evaluation was carried out on the Enhanced CK dataset. Overall, the results indicate that CNN is very promising and surpassed all other methods

1 | INTRODUCTION

As humans, we are particularly gifted at recognizing other people and inferring their mental states from even a cursory glance at their faces – in fact, even young children can recognize happiness and emulate smiles. If Artificial Intelligence and Robotic systems are to spread wider in society, they will need to be able to interact with people appropriately by recognizing and taking into account their mood and state of mind: to detect, understand, and react to the various affective states. For example, the exact same words can carry very different meanings if spoken in anger, annoyance, amusement or anxiety such as a through feelings of rage, resentment, bitterness, discontent or irritation. One essential difficulty of this task is the large range of differences between people's faces and how they show emotion. In addition, the timing, the speed and the relative duration, and the appearance of the various facial action activation might differ from spontaneous behaviour. Figure 1 shows how the temporal dynamics importance of AUs has a crucial impact on the real meaning of facial expression and distinguishes between posed and spontaneously occurring expressions.

The human face is able to display an assortment of facial expressions. Facial expression is one of the most informative key channels of non-verbal communication by cogent natural way and concerns the facial atomic muscle component movements. The Facial Action Coding System (FACS) is the most comprehensive system that precisely describes the basic facial expression movements by encoding the configuration of AU or multiple AUs in terms of facial atomic activation muscle actions. In a muscle-based approach, FACS defines 46 action units assumed as the smallest fundamental measurement of visible discernible blocks of facial movements [1–3]. Further, this system supports mapping from facial appearance changes to emotion space. In the past, proposed approaches to automatic facial expression analysis were mostly limited to basic emotion categories (happiness, sadness, surprise, fear, anger, and disgust). However, it is not certain whether all facial expressions can be classified under those six basic emotion categories [4]:

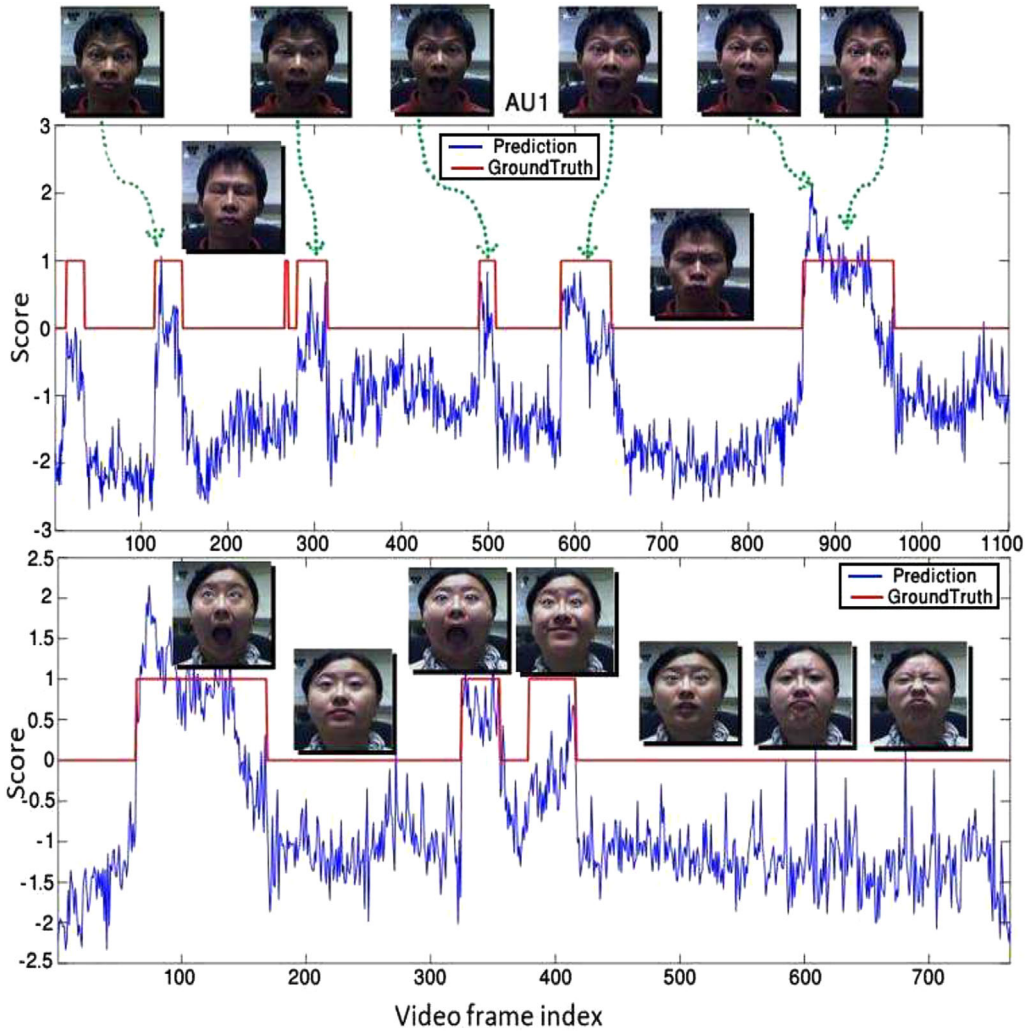

FIGURE 1 An example of using temporal information. The figure represents a continuous scoring prediction, detection of AU1 on the first half part of the sequence (subject 1), and on the second half part of the sequence (subject 2) in which we used the feature vector from the enhanced CK dataset for training and for testing a feature vector which included a sequence of two videos with two subjects. Each one consists of 900 frames from the ISL Facial Expression dataset using LPQTOP dynamic descriptor

people can often show a mixture of emotional expressions. Furthermore, pure facial expressions are rarely elicited. Yet to date, psychological research on this topic remains scarce. Moreover, from a technical standpoint, detecting real-time facial expression already presents a difficult challenge in computer vision due to the level and ambiguity of the variability, the subtlety, and the complexity in its appearance and subjects can be extremely dynamic in their pose. Facial expression analysis refers to computer applications that are designed to automatically recognize facial feature changes using visual information. Facial changes can be identified as facial action units or prototypic emotional expressions depending on whether the temporal information is used. This involves many sub-problems which are not yet fully solved: detection of an image segment as a face, extraction of information from the facial region, and classification of facial AUs. Ideally, the typical structure of automatic facial AU recognition processes consists of multiple steps, in three main stages:

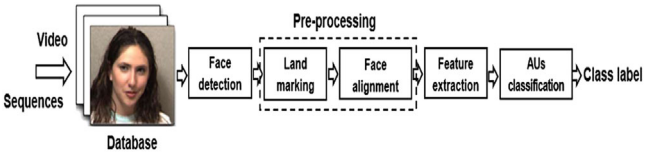

FIGURE 2 The ideal proposed system

detection of facial regions/alignment and tracking, facial feature extraction, and AU classification [7] (Figure 2).

Face detection typically serves as the first initial step across facial analysis pipelines. Arguably, a popular strategy for finding a face bounding box uses the classic real-time Viola–Jones method. There are many available techniques for face detection and numerous tools exist in the field, for example, the dlib, Seetaface, FaceReader2, Av+EC2015, Emotient1, IntraFace,

and NVSIO3 [8, 9]. Face tracking is another aspect of facial expression analysis which can often be a consequence of face detection. Tracking means realizing the face in that frame of sequence is identical to the same face in the last frame of the sequence. Face landmarking is denoted as the detection and localization of certain key characteristic points on the face. These points are used to represent the information required to classify an individual and to determine local patches to extract features for AUs prediction. Landmarks are represented by the centres and corners of the eyes, nostrils and mouth corners, ear lobes, nose tips, the eyebrow arcs, cheeks, and chin. These details are called fiducial landmarks in the face processing literature. Moreover, the purpose of face alignment is to locate facial landmarks automatically and to map the rectified face image into the same canonical pose view (typically, the front view) which is important for some tasks, such as face tracking, security monitoring, facial expression recognition, and 3D face modelling [10]. After the face is detected, in this step, feature extraction methods are used to extract a feature vector (features) that is fed into a classification system. Feature extraction techniques can be divided according to whether they focus on motion or deformation of faces and facial features [5]. Classification techniques are conducted for supervised learning algorithms such as Euclidean distance classifier, nearest neighbour classifier, Fisher face [5], neural networks, discriminant analysis, support vector machines (SVM), and hidden Markov models (HMMs) [6]. Classification and predictions of AUs are the output of the system and the final step in the pipeline.

The novelty of this work starts by proposing a benchmark of dynamic versus static methods for facial expression recognition. The potential advantages of this work are to design an automated system that is capable of recognizing and estimating the emotions of different individual's feelings in real-time from live broadcast footage. The proposed system can significantly advance the existing work from different aspects and will extend the state-of-the-art knowledge boundaries by looking at how emotional cues can be learnt and recognized by discovering temporal changes in facial appearance and how such patterns learnt on test subjects can be generalized for applications to new individuals. Modelling and recognizing people's emotion from their faces, achieved by recognizing action units (AUs) is a challenging computer vision problem. Emotions are usually described in terms of individual action units (AU), the atomic components of the facial expression of emotions. In real-world applications, machines that interact with people need strong facial expression recognition. This recognition is seen to hold advantages for varied applications in affective computing, advanced human–computer interaction, security, stress and depression analysis, robotic systems, and machine learning.

Our aim is to address three main complementary aspects: the problem of modelling AU target activation detection, and then, to discover the underlying temporal variation phases in a sequence using supervised and unsupervised methods which highlight and compare the exciting feature extraction representations on both static and dynamic data, which confer the importance of fusing more than one deep architecture. The proposed methods were evaluated by the third aspect: comparing the continuous scoring predictions by acquiring a best match between the predictions and the ground truths. We demonstrated that both methods (static and dynamic) can compete with the state-of-the-art available methods and the results were promising when tested on the available enhanced Cohn–Kanade dataset and the achieved results illustrate the effectiveness of the proposed methods.

This paper is organized as follows: after this introduction in Section 1, Section 2 briefly gives an up to date review along the topic challenges and summarizes recent work and developments in this domain. The methodology of the feature extraction methods proposed in both categories, static and dynamic, are presented together with the proposed hybrid recognition architecture, detailed in Section 3, which also discusses the experimental settings and gives the results in Section 4, respectively. The conclusions are provided with possible future directions in Section 5.

2 | BACKGROUND

Recognizing AUs automatically from videos is undoubtedly a complex and challenging task. There are several obstacles associated with facial expression recognition which can be traced to many confounding factors which can significantly affect the system performance, and the accuracy of the level of classification [4]. This includes the following: illumination is one of the biggest difficulties for automated facial expression recognition systems. Illumination varies owing to different levels of skin reflection, lustre from eyes, teeth, and camera [13]. Non-frontal pose variation (in a plane, or out of plane rotation) and face misalignment in invariant head movement is a significant research problem found in unconstrained face recognition systems because of the 3D dynamic nature of a facial action [13]. This includes various identities across subjects such as babies, children, youngsters, adults, and elders. Subtle or large individual attribute differences between people's faces occur in key facial features such as intensity, appearance, shape, and conformation to the same facial expression. Imbalanced data with a scarce and limited AU image coded data annotation, according to the lack of adequately FACs coded dataset, represents a major issue impeding progress in the field. Another challenge is that facial AU events can occur in very different time scales [11]. In real time, in most cases, certain positive examples of AUs are minimal, owing to the rarity of becoming activated due to natural facial expression (such as AU9 or AU20). This has to be taken into consideration to avoid 'overfitting on the training data' [12]. Finally, other factors are adversely susceptible such as registration errors, low intensity of facial expressions, noise and occlusions, time delay, age progression, face size, mood and behaviour, scale and orientation, motion blur, gender, ethnicity, facial hair, recording environment, permanent furrows, decorations, accessories and skin marks, make-up, glasses, piercings, tattoos, beards and scars which can either occlude or obscure the face [13, 14, 17] [18, 19]. Facial AU recognition holds a vast number of potential applications from computer vision, surveillance, facial animation, tiered detection, health care,

psychological inquiry, social robotics, pain assessment, driver safety system, behaviour interpretation science, the orientation of the degree of attention of characters in videos, interactive video games, intelligent transportation, online avatars mimicking humans, feelings detection, early detection of numerous diseases, and human–computer interaction along with virtual reality [20, 21]. In general, facial AU recognition methods can be divided into three categories. Frame level-based approaches detect and evaluate AU occurrences (facial texture changes such as bulges and wrinkles) in each frame independently using appearance or geometric feature extraction methods, combined with binary classifiers such as SVM or Adaboost [22]. While all the methods try to find landmarks, features location information, or the geometry of the facial shape components signifies geometric features. Segment-level approaches use temporal dynamics in video sequences to detect AU from a set of temporally contiguous frames. Temporal phase modelling algorithms (transition detection) seek to discover constituent temporal segments: neutral, onset, apex, and offset in the event episode [1, 11, 23–25]. In the past, to date, many approaches adopted various conventional hand crafted feature representations for facial AU recognition, that can be broadly divided into appearance, geometric, dynamic, and fusion such as local binary patterns (LBP) and the family of descriptors of engineered representations: LBP histograms from three orthogonal planes (LBP-TOP), local Gabor binary patterns from three orthogonal planes (LGBPTOP), Gabor motion energy, histograms of local phase quantization (LPQ), and their spatial/temporal extensions merits: local phase quantization from three orthogonal planes (LPQTOP) [26], edge orientation histogram (EOH) [27], facial landmarks, histogram of optical flow [11, 20], speed up robust features (SURF), Principle Component Analysis, Gabor wavelets, sparse learning, discrete cosine transform (DCT), histogram of oriented gradients (HOG), 3D HOG [28], pyramid histogram of oriented gradients (PHOG) [29], DAISY/scale invariant feature transform (SIFT) descriptors [30, 31], 3D SIFT [32], Non-negative matrix factorization, and motion history images (MHI) [33]. However, the aforementioned methods rely on specific problems under certain uses. Intuitively, while facial actions express themselves over a time span, a dynamic pattern information captures the trajectory changes of current state, and past state in a time space volume [34]. On the other hand, frame-based methods are faster and easier to implement. However, static methods are very restricted in detecting affective expressive actions in real time, conveying less important information and neglecting to handle the latent temporal variations among consecutive frames of the sequence [20]. On the other hand, some AUs can be recognized using static features only, and also the remain dynamic features are important; for example, the only lone difference between AU43 and AU45 lies in the area of temporal duration of eye closure. Nevertheless, a static image can often still provide enough beneficial information for AUs recognition [1]. The question is whether the detection of the occurrence of target AUs needs the modelling of the entire sequences, or whether a single frame is sufficient.

FIGURE 3 The rules used to represent an uncontrollable rage expression by the activation of AU1, AU2, AU5, AU6, AU9, AU10, AU25, AU26, and AU27

A plethora of published work on dynamic facial expression analysis has concentrated on incorporating the temporal relations of the frame order continuity in a sequence to improve the performance of video prediction. Previous studies which used a group of heuristic rules-based per AU with facial landmark positions [1], such as Figure 3, represented an uncontrollable rage expression from the GEMEP-FERA dataset using some rules for mapping AUs to emotions by the activation of AU1, AU2, AU5, AU6, AU9, AU10, AU25, AU26, and AU27. Discriminative graph-based methods such as variants of dynamic Bayesian network (DBN) are probabilistic graphical models that can learn the full conditional joint probability of temporal cues for facial actions [22], such as Conditional Random Fields, Latent Dynamic Conditional Random Fields [24], the Kernel Conditional Ordinal Random Field, and Hidden Conditional Random Fields for action unit estimation. Hidden Markov chain transition models are used to encode temporal persistence and the likelihood of label transitions throughout the sequence [17]. Weakly supervised learning such as Multiple Instance Learning are proposed to deal with incomplete labels. A semi-supervised learning approach can be effective in recognizing all the positive samples of annotated data with potentially advantageous unlabelled data [35]. Segment-based classifiers use a bag of temporal words to represent the segments. For unsupervised approaches; Sequence-based clustering algorithms are used to group events of similar characteristics. Slow Feature Analysis describes a latent space time variation that correlates with the AU temporal segments [36]. An unsupervised Branch-and-Bound framework is used to force synchrony correlated facial actions in an unannotated sequence [8].

On top of that, more recent work using Deep Convolution Neural Networks, involving robust accurate learning for more discriminative feature extraction from raw pixel image data, has triumphed over traditional methods. This is due to their exceptional ability of reporting improved results stemming from desired characteristic representations which result in high performance to expedite the process of training and testing at

very low power consumption in many computer vision tasks, for example, object detection, facial expression recognition, image classification, and scene understanding [2]. One of the major limitations of conventional CNN is that impartially extracted spatial relations of the facial components cannot consider the temporal variation relations [11, 37]. An alternative is to utilize deep neural networks, particularly CNN as a feature extraction way, and then implement an extra classifier, for example, SVM or RF to get the optimal image representations. A recent breakthrough of deep hybrid approaches fusing a CNN and Long Short-Term Memory was developed for combining high-level spatial features while preserving temporal dependencies simultaneously [37, 38].

3 | METHODOLOGY

3.1 | Local Binary Patterns

LBP and its extensions were originally proposed for grey scale invariant image texture analysis. Since then, it has proved to be a very efficient feature descriptor used in many applications because of its computational simplicity and discriminating power for texture classification in real world complex settings. It also remains robust to monotonic greyscale changes, in addition to its sensitivity to local structure tolerance to variations in face alignment [39], though it is not robust to rotations and is prone to noise. In practice, an 8-bits binary pattern (LBP code) response of a pixel is computed, in other words, the image labels are made by comparing and thresholding the value of a central pixel intensity with the intensity of all the local pixels in the neighbourhood. If the intensity of the central pixel is larger or equal to its neighbour's, it is encoded by one, or otherwise zero [40]. Later on, in the aforementioned process each bin will correspond to one of the different possible binary patterns and produce a flow of binary numbers with eight surrounding pixels which will end up with 256 possible combinations of LBP dimensional descriptor. A review of LBP descriptor can be found in [1].

3.2 | Local Phase Quantization

The local phase quantization (LPQ) operator is a static local appearance, texture descriptor using the 2D Short-Term Fourier Transform Phase (STFT) on local image windows neighbourhoods [15], was first suggested as a texture descriptor by Ojansivu and Heikkila [16]. Both LBP and LPQ have been applied successfully for AU recognition and are resistant to image blur. LPQ depends on the blur invariance possession of the Fourier phase spectrum. In LPQ we used only four complex coefficients related to 2D frequencies. The phase information, the real and the imaginary part for each pixel position in the Fourier coefficient is calculated through a rectangular M-by-M neighbourhood and is recorded by keeping the signs of the real and imaginary parts of each component [17]. As a result, we get a 256-dimensional feature vector from 8-bit binary coding coefficients, represented as integers.

3.3 | LPQTOP

The LPQTOP descriptor [26] is an extension of the basic LPQ operator to the time domain where the LPQ features are extracted autonomously from three orthogonal slices, denoted by *x-y*, *x-t*, and *y-t*, respectively [9]. The main advantages of the LPQTOP descriptor are robustness against image transformations such as rotation, insensitivity to illumination variations, computational simplicity, and multi-resolution analysis. The LPQTOP dynamic texture descriptor was originally introduced to extract the latent temporal information clues (learn feature representation from video volume), demonstrating facial appearance changes occurring in facial AUs, in terms of expressing temporal segments of facial AUs [1]. On the other hand, LPQTOP encompasses texture analysis and combines static local appearance with shape attribute features (*x-y* plane provides texture spatial domain) and motion change features (*x-t*, and *y-t* planes provide the temporal information domain), in three directions (*x-y*, *x-t*, *y-t*) to encode the phase transition information per image position for each space and time volume, exhibited in facial expressions [9], Figure 4. For more details see ref.[1].

The consequence resulting from binary patterns is stacked for the three orthogonal planes and is concatenated in a single histogram [9]. In the end, we got 768 bins = (256× 3) LPQ-TOP features extracted per spatial–temporal volume containing 3, 5, or 7 s window frames. In our experiment, all the images of Cohn–Kanade are in frontal view and therefore it is not necessary to consider in plane head movement. We split the cropped face region of the input frame of size 256 × 256 pixels in to 10 × 10, 5 × 5 ,7 × 7 blocks separately with a different frame rate each sequence. The optimal size of temporal windows was investigated in dynamic descriptors as Figure 5 explains: the area under the ROC curves (AUC) for AUs activation detection using LPQ-TOP descriptor with two classifiers (SVM and RF) based on different parameters. Lastly, SVM and random forests were used as binary classifiers for predicting the occurrence of AUs.

3.4 | Non-linear-slow feature analysis

Facial AUs temporal dynamics analysis can be modelled using the non-linear Slow Feature Analysis method. The SFA was first investigated as an unsupervised learning approach for describing the most slowly time-varying visual facial sequences latent space features of rapidly temporal varying signals that grasp time dependencies, ranked by their continuous temporal consistency. More precisely, it aims to minimize the temporal variance of the approximated first order time derivative of the input signal which seeks uncorrelated projections [41, 42]. However, 'Despite its interesting theoretical aspects, the practical applicability of purely unsupervised learning is not clear' [17, 36].

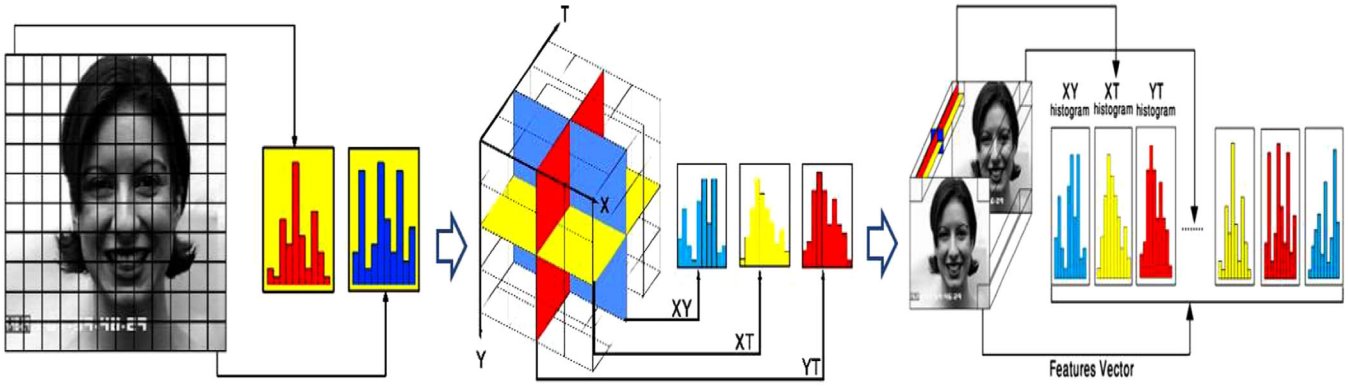

FIGURE 4 LPQTOP descriptor: Block features extracted from all the three planes are histogram concatenated to create a feature vector which represents the whole sequence

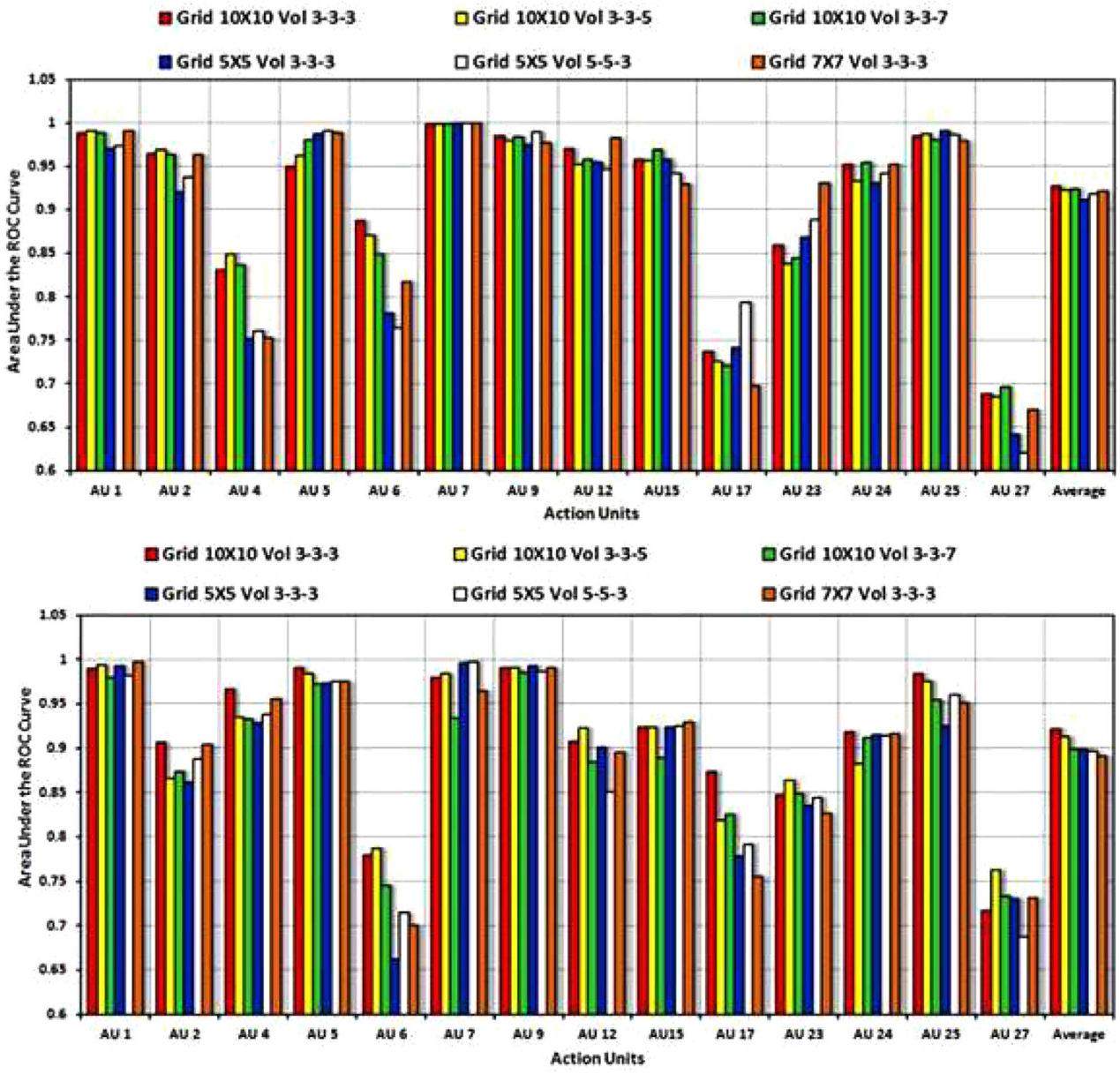

FIGURE 5 AU activation detection using LPQTOP descriptor with two classifiers (SVM& RF) based on different parameters

As of our knowledge, until today, there is limited interesting work focusing on revealing the dynamics of AUs using nonlinear SFA in an unsupervised way regarding its ability to discover the temporal phases of AUs and their constituent temporal segments (onset, apex, offset) [42]. To do so, we applied the method presented by [41], and this can be accomplished by using an expansion function to extend the input signal data nonlinearly, reducing the dimensionality and track by linear SFA.

3.5 | Long short-term memory

The Long Short-Term Memory (LSTM) is a special type of temporal fusion densely connected recurrent neural network modules proposed by Hochreiter and Schmidhuber [43] to solve the problem of vanishing/exploding gradients encountered by a recurrent neural network. It is embedded to learn long-short dependencies [43]. Notably, LSTM has proven to memorize information for a long time and store context temporal actions, including the previous feature's time step and current states with a time lag [34], in contrast with other classifiers such as HMMs. Wei et al [2], assert that having the former state of a facial action expression can absolutely improve the detection of AUs. Recently, LSTMs were used for sequence processing problems with clear contexts, for example, audio analysis, speech recognition, image caption generation, video captioning, forex forecasting, video action recognition [2], and signature verification [34, 44]. It likewise possesses two advantages: LSTM is fine-tuned end to end with other models and it supports both fixed and arbitrary length inputs or outputs. A common LSTM architecture is a chain-like figure of a repeated design of four units: cell, input gate, output gate, and a forget gate [37, 45].

3.6 | The AlexNet CNN model

Used as a pre-trained feature extraction network, this was designed by the Super Vision group of Alex Krichevsky [46], which mainly consists of 13 convolution layers followed by 5 max-pooling layers and Rectified Linear Units (ReLU) for the non-linearity functions to reduce training time, with 3 fully connected layers at the top of the layer stack which ended up with 1000 ways of softmax. ReLU is used after each convolutional and fully connected layer. It is interesting to notice that AlexNet was the first for introducing dropout layers suggested by [47] to combat the overfitting risks problem, and training time in the fully connected layers, to promote the evolution of huge neural networks. The benefit of data augmentation techniques is employed during training to increase more synthetic additional samples to the network by image transformations and reflections such as rotation, scaling, and flips. Dropout is implemented before the first and the second fully connected layers. This network was competing solely on ImageNet to classify up to 1000 various object classes. The input image size to this network should be 227 × 227 × 3. The CNN model has been pre-trained on the Labelled Faces in the Wild and the YouTube

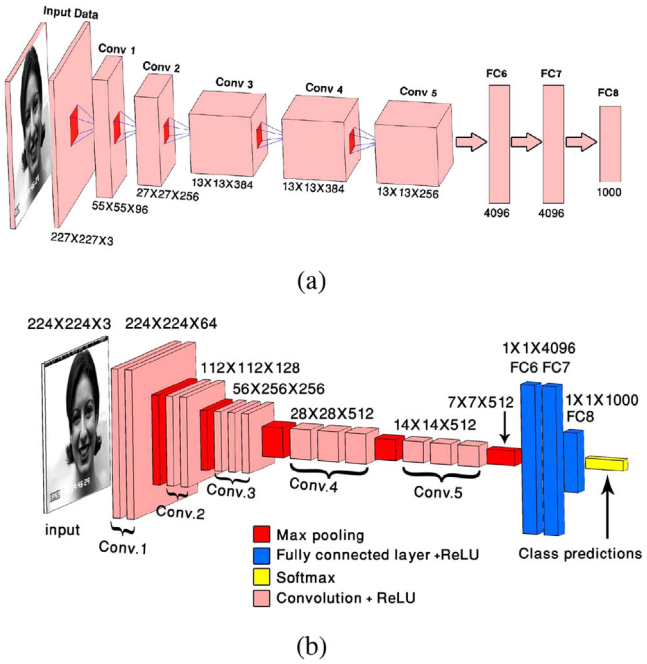

FIGURE 6 Comparison between (a) AlexNet and (b) VGG16

Faces dataset for face recognition [7]; therefore; it will be more suitable for facial expression recognition [2, 11 33, 48].

3.7 | The VGG16 CNN model

Proposed by the VGG team in the ILSVRC 2014 competition, it differs from AlexNet in that it consists of 16 layers which use rich and complex fixed kernel sized filter banks of 3 × 3 (11 × 11 filters in the first layer in AlexNet) for all conventional layers. Using a max pooling of 2 × 2, the number of filters is doubled after each max pooling. After the convolutional layers, it is followed by 3 fully connected layers with 1 × 1 kernel and the output of 512 feature maps. VGG16 is trained on 1.2 million images of size 224 × 224 × 3 belonging to classify 1000 class categories. The two fully connected layers FC6 and FC7 have been used as a feature extraction layer of depth 4096 dimensions to learn the deep rich representations of the given targets. A loss layer softmax is added to the end of the network to adjust the back-propagation error and probabilistic predictions [48]. Figure 6 summarizes the comparisons between the two Convolutional Neural Networks proposed architecture chart.

The authors in [17] point out that for more than 10 years, the academic researchers have held an all-inclusive range of AU labelling databases but in fact only CK and MMI databases are available. For the MMI dataset, the whole sequence is annotated as an active state if the target action unit happens in any frame of the sequence and is classified as a positive of the equivalent video. For instance, AU45 (blink) occurred very quickly in some frames of the video and fundamentally, the entire sequence was labelled as AU45 active, yet the video level annotations for weakly supervised settings (not individual frame level

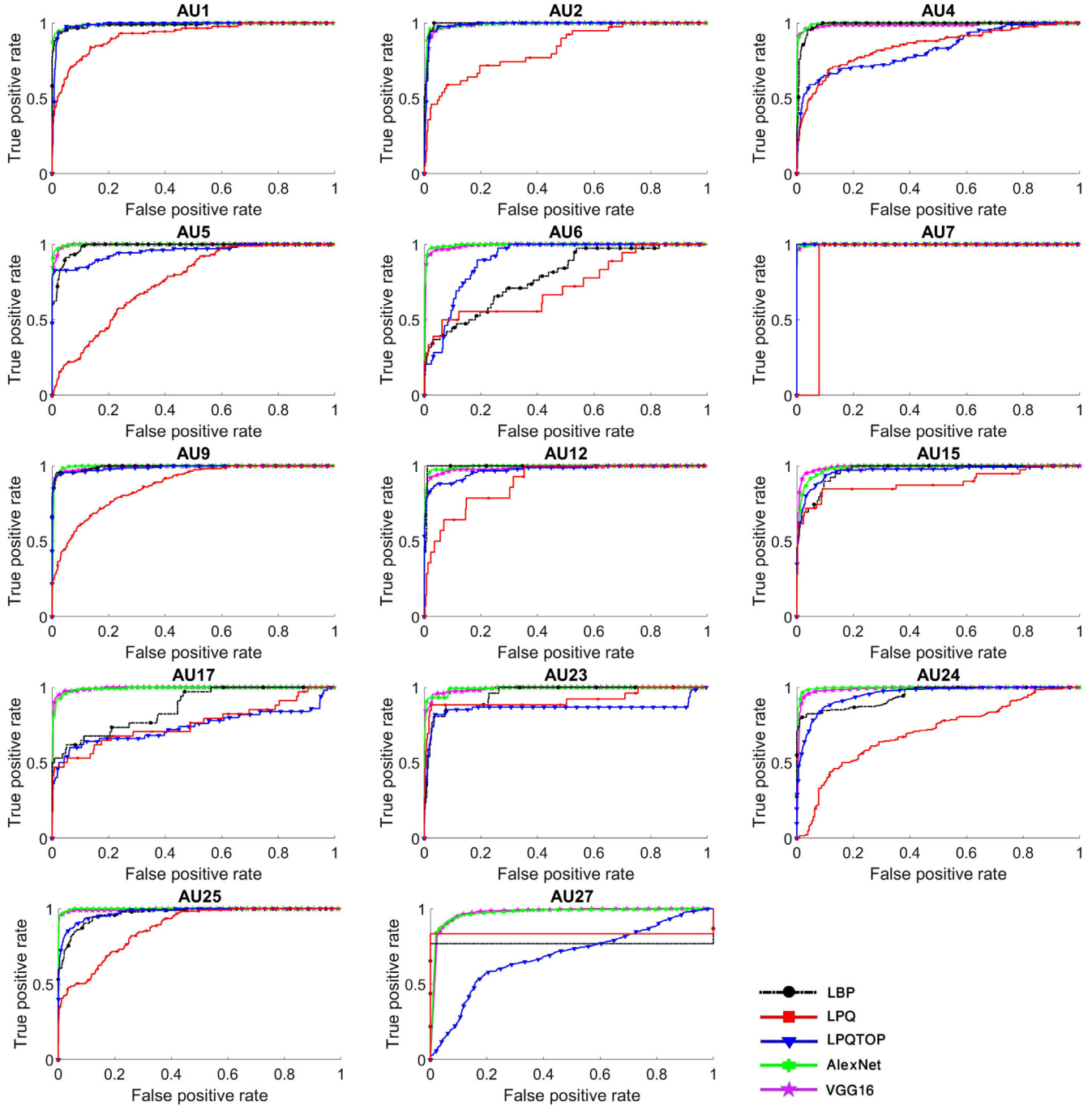

FIGURE 7 Receiver operating curves (ROC) for 14 action units (AU) and 5 dissimilar methods. Each ROC depicts five methods, black: LBP with SVM, red: LPQ with SVM, blue: LPQTOP with SVM, green: AlexNet with SVM, purple: VGG16

annotations), would not have the same truly frame-by-frame basis for AU annotated ground truth. Also, the information on temporal segment detection annotations is concealed for competition, as mentioned in ref. [49]. For these reasons, in our experiments in this paper, we depend on the ISL Enhanced Cohn–Kanade AU-coded Facial Expression Database, in which the Intelligent System lab by Rensselear Polytechnic Institute produced a new AU manual relabelling which counted by the frame-by-frame annotations, which are mostly used for facial action unit recognition [50].

4 | EXPERIMENTAL SETTINGS AND EVALUATION

Three experiments were conducted in this paper on the available enhanced CK dataset comparing features extracted by LBP, LPQ, LPQTOP, AlexNet, and VGG16 for each static image of a video for action unit activation detection, getting hidden insights of underlying temporal variation detection to be investigated by hybrid non-linear SFA(NSFA) + LPQTOP, LPQTOP + LSTM, AlexNet + LSTM, from dynamic sequences.

TABLE 1 AUC values for the first experiment shown in Figure 7

AU	LBP	LPQ	LPQTOP	AlexNet	VGG16
AU1	0.98793	0.92	0.98841	0.99	**0.99157**
AU2	**0.99297**	0.8277	0.9638	0.99022	0.98671
AU4	0.98925	0.84576	0.82542	**0.99605**	0.98685
AU5	0.98431	0.75515	0.95292	**0.99781**	0.99642
AU6	0.78884	0.7279	0.90291	**0.99605**	0.9911
AU7	**1**	0.92124	**1**	0.99913	0.99909
AU9	0.99283	0.8717	0.98857	**0.99436**	0.99181
AU12	**0.99525**	0.89404	0.97069	0.99478	0.98468
AU15	0.96626	0.88785	0.96493	0.98466	**0.99089**
AU17	0.86467	0.75653	0.73945	0.99206	**0.99286**
AU23	0.95694	0.91858	0.86218	**0.99117**	0.99003
AU24	0.9471	0.70181	0.95272	**0.99508**	0.98716
AU25	0.96899	0.8732	0.9772	**0.99824**	0.9949
AU27	0.76856	0.83406	0.68884	0.97135	**0.97286**
Average	0.943136	0.838251	0.912717	**0.992211**	0.989781

TABLE 2 Accuracy values for the first experiment shown in Figure 7

AU	LBP	LPQ	LPQTOP	AlexNet	VGG16
AU1	**0.9835**	0.9381	0.9683	0.9553	0.9620
AU2	**0.9611**	0.8917	0.9525	0.9434	0.9220
AU4	0.9647	0.9389	0.9492	**0.9735**	0.9711
AU5	0.9282	0.8972	**0.9775**	0.9727	0.9600
AU6	**0.9666**	0.9514	0.9292	0.9644	0.9612
AU7	**0.9993**	0.9976	0.9943	0.9873	0.9802
AU9	**0.9692**	0.8278	0.9677	0.9232	0.9505
AU12	**0.9753**	0.9654	0.97061	0.9719	0.9663
AU15	0.9036	0.8984	**0.9434**	0.9327	0.9382
AU17	0.9487	0.9128	0.9621	**0.9778**	0.9767
AU23	0.9569	0.9340	0.9653	**0.9782**	0.9715
AU24	0.8972	0.5321	0.8879	**0.9343**	0.9133
AU25	0.9575	0.8950	0.9620	**0.9873**	0.9838
AU27	**0.9957**	**0.9957**	0.6986	0.9240	0.9244
Average	0.957679	0.898293	0.93815	**0.959**	0.9558

Additionally, comparing scoring prediction detection between the features was extracted by LPQTOP + SVM, LPQTOP + LSTM, and AlexNet on the enhanced CK dataset. For the three experiments the system is contrived to extract two types of features from supervised methods, which are extracted by LBP, LPQ, LPQTOP, AlexNet, Vgg16, LSTM, and unsupervised methods (linear and non-linear SFA, PCA) including hand crafted features represented by LBP, LPQ, LPQTOP, and the learned deep visual features extracted by CNN and LSTM on both static and dynamic data. We limited our evaluation to the problem of AU activation detection because there is no similar database with corresponding ground truths tuned to AU target occurrence detection. The experiments were carried out on the workstation using the Ubuntu Linux system and all the processes of training and testing were accelerated by the NVIDIA GeForce GTX 980 Ti GPUs.

4.1 | First experiment

The aim of the first experiment was to predict the presence or absence of AU occurrence at frame level and to test the performance on the supervised proposed model. On this basis, we extracted the appearance features from both static and dynamic information from the same dataset with respect to frame-by-frame base. Our experiment is conducted by splitting the dataset into 83% of data for training and 17% of data for testing in which we used 7000 frames for the training stage and 1420 frames for testing and the information of test subjects, which was excluded from training and the images of one subject were used in training or testing at the same time. We first located and cropped the face from all the input frame sequences of size 490 × 640 and utilized an adapted Viola–Jones detector. Subsequently, all input frames were resized to be 250 × 250 pixels (this was also done for experiments two and three). In our experiment, all the images of Cohn–Kanade were in front and this eliminated the problem of head pose non-rigid face registration. Next, to encode shape information for LBP, and similarly for LPQ, and LPQTOP, the images were divided into regions to extract LBP, LPQ, and LPQTOP histograms, respectively. The LBP, LPQ, LPQTOP features extracted from each block are stacked into a single feature histogram. Then, the resulting final histogram is used as a feature vector to represent facial image. For LBP a region size of 32 × 32 is used. That is, the face image is divided into 10 × 10 blocks. Normalisation was done for the obtained histograms in the range between [-1 : 1], and then we get a feature vector of 256 dimensions. For LPQ a local window of size equal to 7 and 4×4 blocks is the optimal choice. For the LPQTOP spatial/temporal descriptor the important parameters are temporal window length (volume size) and spatial block grid size. The average performance is evaluated in a subject independent manner using different parameters. So, the experiment is carried out to find the optimal length and width of the histogram block: ((grid 10 × 10 Vol 3-3-3), (grid 10 × 10 vol3-3-5), (grid 10 × 10 Vol 3-3-7), (grid 5 × 5 Vol 3-3-3), (grid 5 × 5 Vol 5-5-3), (grid 7 × 7 Vol 3-3-3)). Next, the typical linear kernel SVM and RF classifiers are trained separately to detect the occurrence of 14 AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU12, AU15, AU17, AU23, AU24, AU25, AU27) irrespective of the absences or the presence of other AUs. In our case, AUC is our performance metric on a frame-by-frame base and is a better ranking-based measure than other metrics, especially in a balanced class binary classification context [8]. In Figure 7, the prominent LBP is clearly superior to LPQ for most action units; similarly, we present the increased relative performance gained by comparing the performance of LBP and LPQ with dynamic features of LPQTOP respectively.

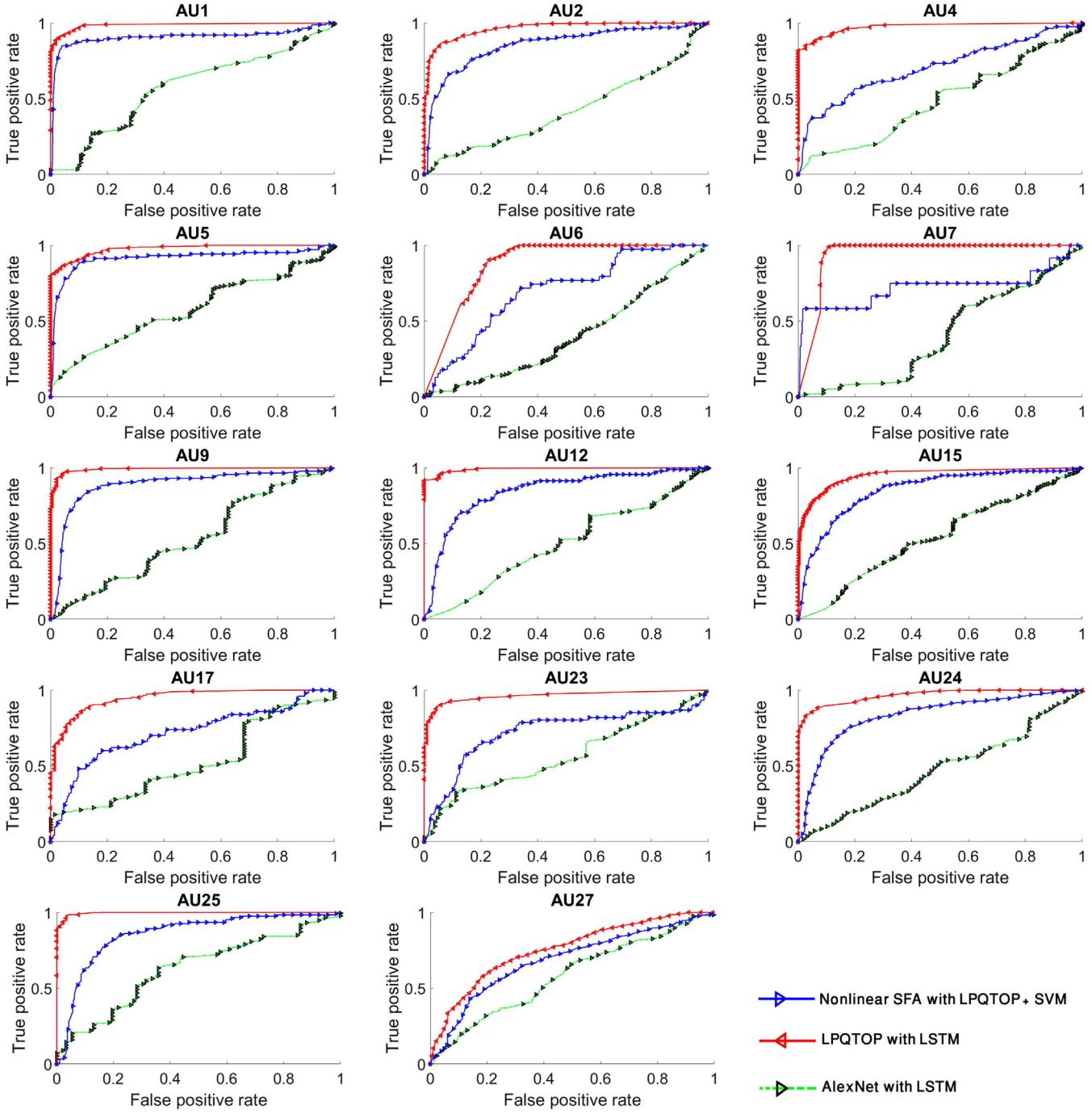

FIGURE 8 Receiver operating curves (ROC) for 14 action units (AU) and 3 dissimilar methods. Each ROC depicts three methods, red: LPQTOP with LSTM , blue: Non-linear SFA with LPQTOP, green: AlexNet with LSTM

It was reported by [1] and [51] that the LPQTOP dynamic appearance descriptor has been presented as superior for the AU activation detection problem and AUs temporal segments recognition. In addition to that, in [51], it was shown that LPQ achieves higher performance than LBP while [52] concluded that the fixed length window is not appropriate for changing facial actions speed. Our experiment showed that LBP clearly overcomes LPQ, and LPQTOP. We also selected two popular pre-trained CNN architecture models: the AlexNet and VGG16 to extract the probability predictions of the cropped faces, in the same way for spatial facial feature representation. Using a pre-trained network model can attain very good foremost parameters to expedite the operation of training and testing. We observed that the heavy computation burden and the time elapsed of extracting the features using the activations from the fc6 and fc7 layers as spatial facial learned features is being less

TABLE 3 AUC values for the second experiment shown in Figure 8

AU	NSFA, LPQTOP	LSTM, LPQTOP	AlexNet, LSTM
AU1	0.90607	**0.98616**	0.57259
AU2	0.85748	**0.96482**	0.42819
AU4	0.70504	**0.97462**	0.48538
AU5	0.91509	**0.97646**	0.57634
AU6	0.69804	**0.87946**	0.40306
AU7	0.72461	**0.94159**	0.40704
AU9	0.88865	**0.99087**	0.53262
AU12	0.84715	**0.99183**	0.49982
AU15	0.85302	**0.95114**	0.54175
AU17	0.71731	**0.94882**	0.52354
AU23	0.7245	**0.96361**	0.56597
AU24	0.83395	**0.9655**	0.47486
AU25	0.85649	**0.99641**	0.62957
AU27	0.68757	**0.75688**	0.58425
Average	0.801069	**0.949155**	0.51607

and reduced significantly. As illustrated in Figure 7, Tables 1 and 2, the best performing features for this task is the AlexNet which vastly outperforms all others in both training and testing evaluation with an average score of 0.992211 for all the AUs, while the second best score was 0.989781 achieved by the VGG16 without any need to increase auxiliary GPU units. Our results demonstrate that our models were adept at learning the supervised task; we were therefore able to avoid any risk of overfitting.

4.2 | Second experiment

For the second experiment, to provide a better inspection of the performance of the tested methods for modelling the temporal facial behaviours and to test the hypothesis of dynamic advantages, as depicted in Figure 8, and Table 3, we employed a new integration feature strategy to preserve the temporal order dependency relations, present in the different frames of the sequences, by feeding the feature vector extracted by LPQ-TOP and jointly trained them using the LSTM model to classify and yield a prediction of per-frame for 14 AUs. This could also show the overall AU activation detection which could benefit best capture from the deep dynamic appearance features construction. The proposed LSTM architecture was trained for 150 epoch iterations on mini-batches of 25 samples. Next, the output scores of CNNs, especially AlexNet and LSTMs, were further aggregated into an averaging fusion network in which both are spatially and temporally deep to train CNN and LSTM simultaneously in an end to end framework, accelerating improved future predictions throughout the two networks. To this end, the main reason we did not endeavour to establish a relative comparative evaluation baseline of this experiment, with

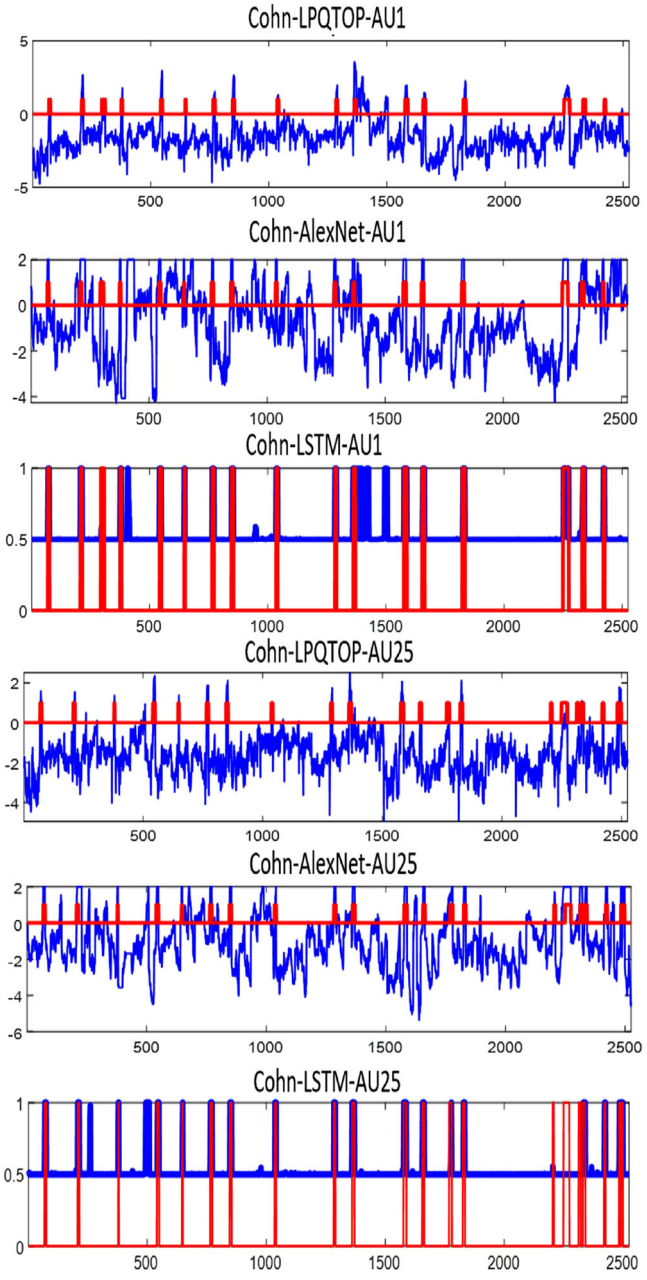

FIGURE 9 Continuous scoring predictions between the three methods for AU1 and AU25

the state-of-the-art deep facial action unit recognition methods, was because there was no existing research paper that could help as the baseline ground truth for all the AUC results (most of the paper use only some of the action units and not all of them), and the majority of them use an F1 measure for metric evaluation. Between them, the non-linear Slow Feature Analysis method was applied as unsupervised learning on also the LPQ-TOP feature vector, after alleviating the dimensionality of the feature vector using Principle Component Analysis which preserved 85% of explained variability leading to a reduced basis of 1,391 dimensions followed by linear Slow Feature Analysis. The first identified latent feature which we obtained corresponded with the most slowly varying one, since non-linear SFA orders

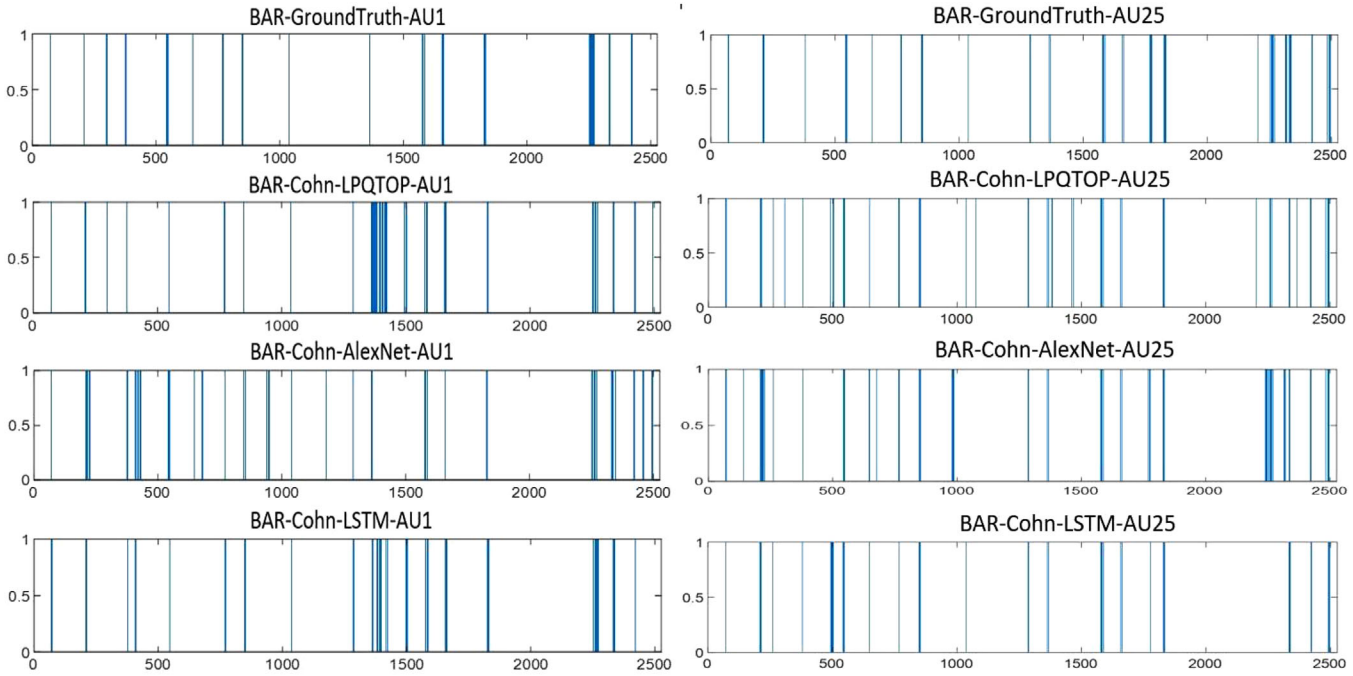

FIGURE 10 Bars of continuous scoring predictions detection using a threshold for best matching between the three methods

TABLE 4 Comparison of the AUC values with the state-of-the-art methods

AUs	A	B	C	D	E	F	G
AU1	0.94	0.95	0.889	0.98793	0.99	**0.99157**	0.98616
AU2	0.97	0.97	0.875	**0.99297**	0.99022	0.98671	0.96482
AU4	0.86	0.89	0.811	0.98925	**0.99605**	0.98685	0.97462
AU5	0.95	0.97	—	0.98431	**0.99781**	0.99642	0.97646
AU6	0.92	0.94	0.94	0.78884	**0.99605**	0.9911	0.87946
AU7	0.78	0.81	0.916	**1**	0.99913	0.99909	0.94159
AU9	0.98	0.98	—	0.99283	**0.99436**	0.99181	0.99087
AU12	0.91	0.93	0.928	**0.99525**	0.99478	0.98468	0.99183
AU15	0.80	0.83	0.982	0.96626	0.98466	**0.99089**	0.95114
AU17	0.84	0.86	0.96	0.86467	0.99206	**0.99286**	0.94882
AU23	0.91	0.92	—	0.95694	**0.99117**	0.99003	0.96361
AU24	—	—	—	0.9471	**0.99508**	0.98716	0.9655
AU25	0.97	0.97	—	0.96899	**0.99824**	0.9949	0.99641
AU27	**1.00**	**1.00**	—	0.76856	0.97135	0.97286	0.75688
Average	0.899	0.915	0.913	0.943136	**0.992211**	0.989781	0.949155

the derived latent variables by their temporal slowness. The performance analysis of this model performs well for detecting the temporal information of AUs. As we pointed out earlier, the feasible application of unsupervised learning in a pure manner is ambiguous; therefore, learning a high-level representation from dynamic textures directly by SFA is not practical because of the curse of dimensionality. We demonstrated that it is possible to use non-linear SFA for accurately discovering the dynamic of facial action units.

4.3 | Third experiment

To assess the ability for maximum expression of the desired target AUs and the classification quality of the described methods, for the third experiment, we compared three types of validation matching the predicted scores which represented the probability of activation for three methods, and the AUC was calculated for AU1 (LPQTOP + SVM AUC = 0.9790, LPQ-TOP + LSTM AUC = 0.9733, AlexNet + SVM AUC = 0.9646)

TABLE 5 Comparison of the accuracy with the state-of-the-art approaches

Methods	Accuracy%
Baseline [57]	0.833
3DCNN-DAP [64]	0.88
3D Shape [54]	0.868
ExpNet [59]	0.612
ITBN [60]	0.863
DTAN + DTGN [61]	0.95
H-CRF [65]	0.88
NMF + ℓ_1 norm [67]	0.924
Spatio-temporal CNN [55]	0.737
DLBP + LL + TFP [62]	0.929
HOG-TOP, geometric features SVM [63]	0.957
Extreme sparse learning (ESL) [68]	0.927
Gabor [66]	0.938
MSDF + BoW [53]	**0.959**
Ours (AlexNet)	**0.959**

and for AU25 (LPQTOP + SVM AUC = 0.9790, LPQTOP + LSTM AUC = 0.9579, for AlexNet + SVM AUC = 0.9985). Within every frame in the CK dataset, the AUs were annotated as 0 (not present),1 (active), and -1(not sure). For plotting, in order to make the units standardized for comparison, we made every frame with -1 ground truth equal to 0.5, then we had three classes (0, 0.5, 1) for the three methods. As can be observed from Figure 9, the time series plot of AU1 (inner eyebrow), AU25 (lips parted) the detection for each algorithm provides almost different predictions and AU1 and AU25 is a unique feature that can be compared across all the three algorithms making them have the potential to confidently measure AU1 and AU25 accurately. We used 317 of the videos for training and 150 videos for testing. Therefore, in total we used 5891 frames during the training phase and 2529 frames for testing. The representation learned by the proposed methods in Figure 9, was capable of exact prediction of the dynamics of the AU1 and AU25, since it provides more accurate features which in turn matched better with the true label GroundTruth (red line). It seems that the LSTM method is less continuous than the other algorithms. Overall, the performance showed that all the three methods provide better results and are intersected in approximately all the time points that are indicative for detecting and predicting the presence of both AU1 and AU25. To facilitate this analysis further, and to see more accurate matching of the scoring predictions for the three methods, we applied a threshold and drew a bar for each method score in Figure 10. Table 4 shows comparison of the AUC values of the proposed methods (D, LBP; E, AlexNet; F, VGG16; G, LSTM and LPQTOP) with the state-of-the-art approaches (A, SPTS [57]; B, relative AU [19]; C, STM [58]) for AU detection on the extended CK dataset. A comparison of the obtained accuracy was also presented in Table 5, with different state-of-the-art techniques on the extended CK dataset including sparse coding, manifold learning, deep and unsupervised learning.

5 | CONCLUSION AND FUTURE WORK

In this paper, our model was focused on three main essential problems: AU activation detection by confirming the superiority ability of a pre-trained AlexNet that boosts reliably overall average recognition rate and accuracy, which comes up with significant AU prediction scoring improvements and strengthens the requirements of using deep learning, contrary to the traditional hand crafted and engineered features. The second is temporal modelling by testifying that fusing both temporal and temporal features will gain more long-term temporal pattern information. Third, achieving a successful comparison of continuous scoring predictions of AUs activation detection was accomplished which was shown to be efficacious. Our future work will be modelling multiple action unit activation detection as they seemingly appear to build a single display to encode them as an entire facial event for automatic occurrence recognition of an affective state.

ORCID

N. Pugeault https://orcid.org/0000-0002-3455-6280

REFERENCES

1. Jiang, B., et al.: A dynamic appearance descriptor approach to facial actions temporal modeling. IEEE Trans. Cybernetics 44(2), 161–174 (2014)
2. Li, W., et al.: Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In: Computer Vision and Pattern Recognition (CVPR), pp. 6766–6775 (2017)
3. Valstar, M.F., et al.: FERA 2015-second facial expression recognition and analysis challenge. In: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Vol. 6, pp. 1–8 (2015)
4. Shbib, R., Zhou, S.: Facial expression analysis using active shape model. International Journal of Signal Processing, Image Processing and Pattern Recognition 8(1), 9–22 (2015)
5. Bhati, D., Gupta, V.: Survey—A comparative analysis of face recognition technique. Int. J. Eng. Res. General Sci. 3(2), 597–609 (2015)
6. Girard, J.M., et al.: How much training data for facial action unit detection? In: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Vol. 1, pp. 1–8 (2015)
7. Li, X., et al.: Facial action units detection with multi-features and AUs fusion. In: 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), pp. 860–865 (2017)
8. Chu, W.S., et al.: Learning spatial and temporal cues for multi-label facial action unit detection. In: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), pp. 25–32 (2017)
9. He, L., et al.: Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pp. 73-80 (2015)
10. Geng, C., Jiang, X.: Face alignment based on the multi-scale local features. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1517–1520 (2012)
11. Hasani, B., Mahoor, M.H.: Facial expression recognition using enhanced deep 3D convolutional neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2278–2288 (2017)

12. Nicolle, J., et al.: Real-time facial action unit intensity prediction with regularized metric learning. Image Vision Comput. 52, 1–14 (2016)
13. Hatem, H., et al.: A survey of feature base methods for human face detection. International Journal of Control and Automation 8(5), 61–78 (2015)
14. Barr, J.R.: Gallery-free methods for detecting and recognizing people and groups of interest in the wild. University of Notre Dame (2015)
15. Rattani, A., et al.: Evaluation of texture descriptors for automated gender estimation from fingerprints. In: European Conference on Computer Vision, pp. 764–777, Springer, Cham (2014)
16. Ojansivu, V., Heikkila, J.: Blur insensitive texture classification using local phase quantization. In: International Conference on Image and Signal Processing, pp. 236–243, Springer, Berlin, Heidelberg (2008)
17. Martinez, B., et al.: Automatic analysis of facial actions: A survey. IEEE Trans. Affective Computing. 10(3), 325–347 (2017)
18. Han, S., et al.: Incremental boosting convolutional neural network for facial action unit recognition. In: Advances in Neural Information Processing Systems, pp. 109–117 (2016)
19. Khademi, M., Morency, L.P.: Relative facial action unit detection. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1090–1095 (2014)
20. Sariyanidi, E., et al.: Learning bases of activity for facial expression recognition. IEEE Trans. Image Process. 26(4), 1965–1978 (2017)
21. Zhen, Q., et al.: LPQ based static and dynamic modeling of facial expressions in 3D videos. In: Biometric Recognition, pp. 122–129. Springer, Cham (2013)
22. Seckington, M.J.: Using dynamic Bayesian networks for posed versus spontaneous facial expression recognition (2011)
23. Ding, X., et al.: Cascade of tasks for facial expression analysis. Image Vision Comput. 51, 360-48 (2016)
24. Walecki, R., et al.: Variable-state latent conditional random fields for facial expression recognition and action unit detection. In: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Vol. 1, pp. 1–8 (2015)
25. Zafeiriou, L., et al.: Joint unsupervised face alignment and behaviour analysis. In: European Conference on Computer Vision, pp. 167–183.Springer, Cham (2014)
26. Corneanu, C.A., et al.: Survey on RGB, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. IEEE Trans. Pattern Anal. Mach. Intell. 38(8), 1548–1568 (2016)
27. Timotius, I.K., Setyawan, I.: Evaluation of edge orientation histograms in smile detection. In: 6th International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 1–5 (2014)
28. Klaser, A., et al.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC 2008—19th British Machine Vision Conference, pp. 275-1 (2008). British Machine Vision Association
29. Barsoum, E., et al.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 279–283 (2016)
30. Wu, Y., Ji, Q.: Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3400–3408 (2016)
31. Mayya, V., et al.: Automatic facial expression recognition using DCNN. Procedia Comput. Sci. 93, 453–461 (2016)
32. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM international conference on Multimedia, pp. 357–360 (2007)
33. Mollahosseini, A., et al.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1-10 (2016)
34. Vadapalli, H.: Facial action unit recognition from video streams with recurrent neural networks. Int. J. Comput. Appl. 96(19), (2014)
35. Senechal, T., et al.: Facial action unit detection using active learning and an efficient non-linear kernel approximation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 10–18 (2015)
36. Zafeiriou, L., et al.: Deep analysis of facial behavioral dynamics. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1988–1996 (2017)
37. Ko, B.C.: A brief review of facial emotion recognition based on visual information. Sensors 18(2), 401 (2018)
38. Sun, B., et al.: Automatic temporal segment detection via bilateral long short-term memory recurrent neural networks. J. Electron. Imaging 26(2), 020501 (2017)
39. Almaev, T., et al.: Learning to transfer: Transferring latent task structures and its application to person-specific facial action unit detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3774–3782 (2015)
40. Joshi, D.: A brief review of facial expressions recognition system. Asian Journal For Convergence in Technology (AJCT)-UGC Listed 4(I), (2018)
41. Zafeiriou, L., et al.: Probabilistic slow features for behavior analysis. IEEE Trans. Neural Networks Learn. Syst. 27(5), 1034–1048 (2016)
42. Zafeiriou, L., et al.: Learning slow features for behaviour analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2840–2847 (2013)
43. Jaiswal, S., Valstar, M.: Deep learning the dynamic appearance and shape of facial action units. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8 (2016)
44. Fan, Y., et al.: Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 445–450 (2016)
45. Ullah, A., et al.: Action recognition in video sequences using deep Bi-directional LSTM with CNN features. IEEE Access 6, 1155–1166 (2018)
46. Krizhevsky, A., et al.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25, 1097–1105 (2012)
47. Srivastava, N., et al.: Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15(1), 1929–1958 (2014)
48. Jan, A., et al.: Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. IEEE Transactions on Cognitive and Developmental Systems 10(3), 668–680 (2017)
49. Valstar, M.F.: Timing is everything: A spatio-temporal approach to the analysis of facial actions (2008)
50. Qiang, J.: http://www.ecse.rpi.edu/~cvrl/database/database.html/ (2017). Accessed 16 Oct 2017
51. Jiang, B., et al.: Action unit detection using sparse appearance descriptors in space-time video volumes. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), pp. 314–321 (2011)
52. Almaev, T.R., Valstar, M.F.: Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 356–361 (2013)
53. Sikka, K., et al.: Exploring bag of words architectures in the facial expression domain. In: European Conference on Computer Vision, pp. 250–259.Springer, Berlin, Heidelberg (2012)
54. Jeni, L.A., et al.: 3D shape estimation in video sequences provides high precision evaluation of facial expressions. Image Vision Comput. 30(10), 785–795 (2012)
55. Gupta, O., Raviv, D., Raskar, R.: Illumination invariants in deep video expression recognition. Pattern Recognit. 76, 25–35 (2018)
56. Zeng, N., et al.: Facial expression recognition via learning deep sparse autoencoders. Neurocomputing 273, 6430–649 (2018)
57. Lucey, P., et al.: The extended Cohn–Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94–101 (2010)
58. Chu, W.S., et al.: Selective transfer machine for personalized facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. 39(3), 529–545 (2016)
59. Chang, F.J., et al.: ExpNet: Landmark-free, deep, 3D facial expressions. In: 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018), pp. 122–129 (2018)

60. Wang, Z., et al.: Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3422–3429 (2013)
61. Jung, H., et al.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2983–2991 (2015)
62. Ding, Y., et al.: Facial expression recognition from image sequence based on LBP and Taylor expansion. IEEE Access 5, 19409–19419 (2017)
63. Chen, J., et al.: Facial expression recognition in video with multiple feature fusion. IEEE Trans. Affective Comput. 9(1), 38–50 (2016)
64. Liu, M., et al.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: Asian Conference on Computer Vision, pp. 143–157.Springer, Cham (2014)
65. Walecki, R., et al.: Variable-state latent conditional random field models for facial expression analysis. Image Vision Comput. 58, 25–37 (2017)
66. Mahoor, M.H., et al.: Facial action unit recognition with sparse representation. In: Face and Gesture, pp. 336–342 (2011)
67. Zafeiriou, S., Petrou, M.: Sparse representations for facial expressions recognition via l 1 optimization. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 32–39 (2010)
68. Shojaeilangari, S., et al.: Robust representation and recognition of facial emotions using extreme sparse learning. IEEE Trans. Image Process. 24(7), 2140–2152 (2015)

How to cite this article: Alharbawee L, Pugeault N. A benchmark of dynamic versus static methods for facial action unit detection. *J. Eng.* 2021;2021:252–266. https://doi.org/10.1049/tje2.12001