Action Unit detection using sparse appearance descriptors in space-time video volumes

Bihan Jiang, Michel F. Valstar and Maja Pantic

Abstract— **Recently developed appearance descriptors offer the opportunity for efficient and robust facial expression recognition. In this paper we investigate the merits of the family of local binary pattern descriptors for FACS Action-Unit (AU) detection. We compare Local Binary Patterns (LBP) and Local Phase Quantisation (LPQ) for static AU analysis. To encode facial expression dynamics, we extend the purely spatial representation LPQ to a dynamic texture descriptor which we call Local Phase Quantisation from Three Orthogonal Planes (LPQ-TOP), and compare this with the Local Binary Patterns from Three Orthogonal Planes (LBP-TOP). The efficiency of these descriptors is evaluated by a fully automatic AU detection system and tested on posed and spontaneous expression data collected from the MMI and SEMAINE databases. Results show that the systems based on LPQ achieve higher accuracy rate than those using LBP, and that the systems that utilise dynamic appearance descriptors outperform those that use static appearance descriptors. Overall, our proposed LPQ-TOP method outperformed all other tested methods.**

I. INTRODUCTION

Automated analysis of non-verbal behaviour, and especially of expressive facial behaviour, has attracted increasing attention during the past decades. Current human-computer interaction (HCI) designs usually involve traditional interface devices such as the keyboard and mouse and are constructed to emphasise the transmission of explicit messages whilst ignoring implicit information about the user, such as changes in the affective state [21]. However, many facial expressions play important roles in our daily communication. For example, we raise our eye brows to stress the importance of a spoken message. Leveraging these social signals would have many practical applications. For instance, it would open up new possibilities for interacting with computers, and in the field of psychology, this technology could provide a more efficient and objective log of a subject's facial expressions than the psychologists could create unaided.

Deriving an effective facial representation from images is an essential step for successful facial expression recognition [12]. Traditionally the feature extraction approaches may be divided into two streams: geometric feature-based methods and appearance-based methods. Geometric feature based methods employ the geometrical properties of a face such as the positions of facial points relative to each other, the distances between pairs of points or the velocities of separate facial points. For a method using appearance features, the changes in image texture such as those created by wrinkles, bulges, and changes in feature shapes are captured. In [22], a comparison of recognition performance with different types of features shows that the appearance-based features Gabor wavelet coefficients are more powerful than geometric positions. However, this has been disputed by Valstar and Pantic [19]. They have demonstrated that geometric feature-based methods provide similar or better performance than appearance-based approaches in AU recognition. Thus, it is unclear whether an appearance-based or geometric-based feature extraction method is better, or, as seems more likely, each has its own complementary quality, and the two would be best used together.

One limitation of the majority of existing facial expression recognition methods is that they focus on a small set of prototypic emotional facial expressions, specifically fear, sadness, happiness, anger, disgust, and surprise (e.g.,[12], [23], [22]). Yet, these six basic emotion categories form only a subset of the total range of possible facial displays and the categorisation of facial expressions can therefore be forced and unnatural [5]. This should be apparent if one considers which categories boredom and confusion should be placed into, for instance. Moreover, a purely emotion-oriented system would miss other signals sent by the face; for example it would be unable to lip-read. Therefore, we advocate to use the Facial Action Coding System (FACS) [4] instead.

FACS is best known and most commonly used in psychological studies. The coding system defines atomic facial muscle actions called Action Units (AUs). With FACS, every possible facial expression, emotional or otherwise, can be described as a combination of AUs. Currently FACS coding, i.e. applying FACS to videos, is done manually by experts. It takes one hour or more to manually code 100 still images or one minute of videotape in terms of AUs and their temporal segments. An automatic, accurate FACS coding system could vastly increase the amount of data a psychologist could analyse, which would be a breakthrough in the field of psychology. It is therefore no surprise that more and more effort is made to develop a robust real-time AU analysis system. Though much progress has been made, robust real-time facial expression analysis remains difficult due to its subtlety, complexity, and variability [12].

A number of works were reported towards AU detection from image sequences or videos. Tian et al. [13] developed a system to automatically detect 16 AUs from face image sequences using lip tracking, template matching and neural networks. Bartlett et al. [2] employed a user independent

The authors are with the Department of Computing, Imperial College London, UK bi.jiang@imperial.ac.uk, michel.valstar@imperial.ac.uk, m.pantic@imperial.ac.uk

Maja Pantic is also with the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands

fully automatic system for real time recognition of facial actions from the FACS. The system automatically detects frontal faces in the video stream and codes each frame with respect to 20 Action units. Valstar and Pantic [16] presented a system that enables fully automated recognition of 15 AUs from face video based on tracking 20 facial fiducial points. They also proposed a method to fully automatically recognise the four temporal phases – neutral, onset, apex and offset – of facial muscle actions for the first time in the literature [16]. Koelstra et al. [5] proposed a method that detects AUs and their temporal models using Free-Form Deformations and Motion History Images. Motion History Images are a type of dynamic appearance descriptors first used for AU detection by Valstar et al. [19]. A work that utilises the temporal relations between AUs in a facial expression is that by Tong et al. [14]. However, they do not encode an expression's dynamics in the appearance description.

Recently a new group of appearance descriptors has been proposed and sucessfully applied to face recognition, and six basic emotional facial expression recognition. The family of LBP-based detectors (i.e. LBP, LPQ and LBP-TOP) have been reported to attain a better performance in such problems than most existing methods in various comparative studies, with respect to both performance and computational efficiency (e.g.,[12], [23], [10]). Shan et al. [12] recognised seven expressions (including neutral) from different publicly available databases based on the LBP features. In [1], LPQ was applied to face recognition and they reported that the classification accuracy was higher with the new method than with the well-known LBP or Gabor filter bank methods. Zhao et al. [23] applied LBP-TOP to the six basic emotional expressions recognition. It is reported that LBP-TOP clearly outperformed the earlier approaches such as LBP, Gabor and so on. Recently, LBP-TOP has also successfully been applied to human action recognition by Mattivi and Shao [6]. However, as far as we are aware, no work has been done using these features.

This paper investigates the merits of the family of local binary pattern descriptors for FACS Action Unit detection. We compare LBP and LPQ for static AU analysis. For spatio-temporal AU analysis we extend the purely spatial representation of LPQ to a dynamic texture descriptor which is called LPQ-TOP, and compare this with LBP-TOP. Fig. 1 shows an overview of the proposed system.

In order to detect the upper face AUs, we use 9 SVM classifiers, one for each AU, which are trained on a subset of the most informative spatiotemporal features selected by GentleBoost. To extract these appearance features, we first find the face in the input static image or all frames in an image sequence using an adapted version of the Viola and Jones face detector [20]. Next the first frames in image sequences or static face images are registered to remove head rotations and scale variations. Then for the image sequences, a robust automatic image alignment scheme introduced by Tzimiropoulos et al. [15] is employed to align the frames within the sequence. After that, the static image or image sequence is divided into small blocks, and the LBP and LPQ features are extracted from the static image, or if the input is an image sequence the LBP-TOP and LPQ-TOP features are obtained from small space-time video volumes. The histograms from all space-time video volumes are concatenated as a feature vector to represent the corresponding face image or image sequence. The fastest of the systems described in this work, to wit the LBP-based AU detector, is freely available as part of the SEMAINE framework, which can be downloaded from http://semaine.opendfki.de/.

The efficiency of these descriptors is evaluated by an automatic AU detection system and tested on the posed and spontaneous expression data which are collected from the MMI and SEMAINE databases, respectively [18], [7]. Results show that the systems based on LPQ generally achieve higher accuracy rate than LBP system, and that the systems that utilise dynamic appearance descriptors outperform those that use static appearance descriptors. Although the family of LPQ descriptors are more computationally expensive than the LBP's, they attain a higher recognition performance. All in all, the experimental results clearly show that our proposed spatio-temporal descriptor, LPQ-TOP, outperforms all other tested methods for the problem of FACS Action Unit analysis. Finally, we propose a novel way to generate sparse yet complete appearance training data for both static and dynamic AU detection approaches.

Our key contributions are four fold. First, we propose a novel way to generate sparse yet complete appearance training data for both static and dynamic AU detection approaches. Secondly, the well-performing static appearance descriptor LBP is applied to AU detection for the first time in the literature. Third, the LPQ descriptor is employed for the first time to facial expression problems, specifically FACS AU detection. Finally, a novel spatio-temporal appearance descriptor LPQ-TOP is proposed. The experimental results show that our novel descriptor outperforms the three other methods for FACS AU analysis in terms of recognition accuracy.

The remainder of this paper is organised as follows. Section II briefly describes the basic principle of static appearance descriptors LBP and LPQ. LBP-TOP as well as our proposed LPQ-TOP are presented in Section III. The static and dynamic datasets used in our experiments are described in Section IV, while the evaluation procedures and test results are given in Section V. Finally, Section VI provides the conclusions of our research.

II. STATIC LOCAL APPEARANCE DESCRIPTORS

Recognising facial expressions from static images is a more challenging problem than from image sequences, as less information about expressive actions is available. For example, without a neutral reference frame, it is impossible to tell from a still image whether the appearance of the eyebrows indicates a neutral expression, or that the brows are slightly raised. Still, often a single image can provide enough information for AU detection. The static local appearance descriptors LBP and LPQ offer a promising solution to such problems given their proven applicability to face recognition.

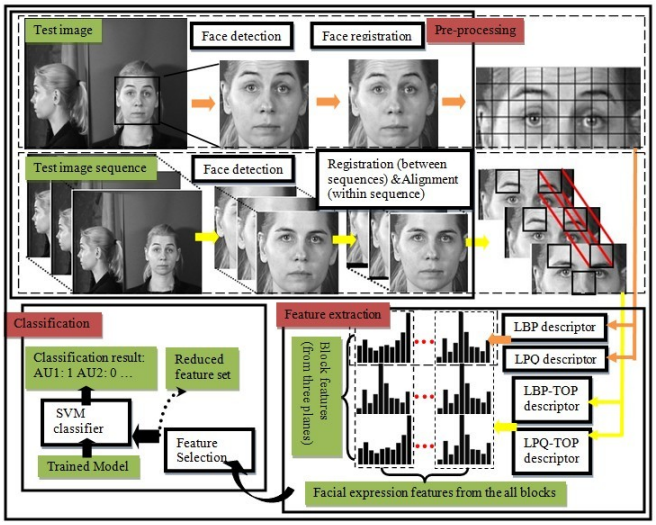

Fig. 1. The outline of our proposed system

In this section, we briefly explain the static appearance descriptors used in this work.

A. Local Binary Patterns

Local Binary Patterns (LBP) were first introduced by Ojala et al. in [8], and proved to be a powerful means of texture description. By thresholding a 3×3 neighbourhood of each pixel with respect to the centre value, the operator labels each pixel. Considering the 8-bit result to be the binary representation of a decimal number, a 256-bin histogram of the LBP labels computed over a region is used as a texture descriptor.

Ojala et al. [9] later extended the basic LBP to a gray-scale and rotation invariant texture operator. They derived an operator for a general case based on a circularly symmetric neighbour set of P members on a circle of radius R, denoted as $\text{LBP}^{\text{riu2}}_{P,R}$. Superscript $''$riu2$''$ reflects the use of rotation invariant uniform patterns (see [9] for details). Parameter P controls the quantisation of the angular space and R determines the spatial resolution of the operator. Bilinear interpolation is used to allow any radius and number of pixels in the neighbourhoods. The rotation invariance is achieved by assigning a unique identifier to each pattern, which is the minimum among the results obtained from performing all the possible circular bit-wise right shifts. A local binary pattern is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular [9]. Using only rotation invariant uniform Local Binary Patterns greatly reduce the length of the feature vector. For example, the number of possible patterns for a neighbourhood of 8 pixels is 256 for the basic LBP while only 59 for LBP^{u2} and 10 for LBP^{riu2}. An early stage experiment was conducted to find the best descriptor and optimal parameters for it. Hence, we adopt the $\text{LBP}^{\text{u2}}_{8,1}$ descriptor in our experiments.

The occurrence of the uniform patterns over a region is recorded by a histogram. After applying the LBP operator

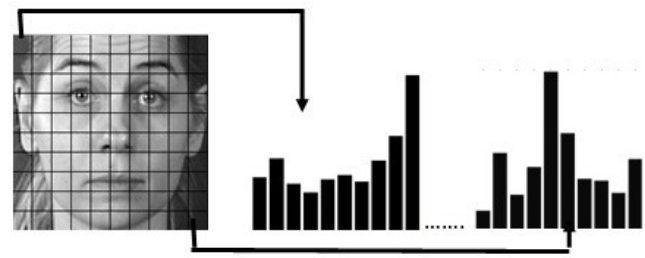

Fig. 2. The concatenated feature vector that extracted from each face block

to an image, a histogram of the labelled image $f(x, y)$ can be defined as

$$H_i = \sum_{x,y} I(f(x, y) = i), \qquad i = 0, ..., n-1 \tag{1}$$

where n is the maximum label number produced by the LBP operator and $I(A)$ is the indicator function, which returns 1 if A is true, and 0 otherwise.

An LBP histogram computed over the whole face image represents only the global distribution of the patterns. To include shape information, the images were divided into regions from which we extract LBP histograms. The LBP features extracted from each region are concatenated into a single, spatially enhanced feature histogram (see Fig.2). The final histogram is used as a feature vector to represent face image. In our experiments, a region size of 20×20 is used. That is, the face image is divided into 10×10 blocks. In our experiments, only the top half of the face is used, as we only detect the upper face AUs.

B. Local Phase Quantisation

The Local Phase Quantisation (LPQ) operator was originally proposed by Ojansivu and Heikkila as a texture descriptor that is robust to image blurring [10]. The descriptor uses local phase information extracted using the 2-D DFT or, more precisely, a short-term Fourier transform (STFT) computed over a rectangular M-by-M neighbourhood N_x at each pixel position $\mathbf{x}$ of the image $f(\mathbf{x})$ defined by

$$F(\mathbf{u}, \mathbf{x}) = \sum_{\mathbf{y} \in N_x} f(\mathbf{x}\text{-}\mathbf{y}) e^{-j2\pi \mathbf{u}^T \mathbf{y}} = \mathbf{w}_\mathbf{u}^T \mathbf{f}_\mathbf{x} \tag{2}$$

where $\mathbf{w}_\mathbf{u}$ is the basis vector of the 2-D DFT at frequency $\mathbf{u}$, and $\mathbf{f}_\mathbf{x}$ is the vector containing all M^2 samples from N_x.

The STFT is efficiently evaluated for all image positions $x \in \{x_1, ..., x_N\}$ using simply 1-D convolutions for the rows and columns successively. The local Fourier coefficients are computed at four frequency points: $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$, and $u_4 = [a, -a]^T$, where a is a sufficiently small scalar ($a = 1$ in our experiments). For each pixel position this results in a vector $F_x = [F(u_1, x), F(u_2, x), F(u_3, x), F(u_4, x)]$. The phase information in the Fourier coefficients is recorded by examining the signs of the real and imaginary parts of each component in F_x. This is done by using a simple scalar quantiser

$$q_j = \begin{cases} 1 & \text{if } g_j \geq 0 \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $g_j(x)$ is the jth component of the vector $G_x = [\mathrm{Re}\{F_x\}, \mathrm{Im}\{F_x\}]$. The resulting eight bit binary coefficients $g_j(x)$ are represented as integers using binary coding:

$$f_{\mathrm{LPQ}}(x) = \sum_{j=1}^{8} q_j 2^{j-1}. \quad (4)$$

As a result, a histogram of these values from all positions is composed and used as a 256-dimensional feature vector in classification.

It can be shown that in quantisation the information is maximally preserved if the samples to be quantised are statistically independent [10]. In practice, the neighbouring pixels are highly correlated in real images, which will lead to dependency between the Fourier coefficients g_j which are quantized in LPQ. Therefore Ojansivu et al. [10] improve LPQ by introducing a simple de-correlation mechanism. Recently Ojansivu et al. [11] presented a rotation invariant extension to the blur insensitive local phase quantisation texture descriptor. For more details, please refer to [10].

Similar to the extraction of LBP histograms mentioned above, the LPQ histograms were extracted independently from non-overlapping rectangular regions and then concatenated to build a spatially enhanced description of the face. For example, an image divided into $m \times n$ blocks will produce a feature vector with dimension of $256 \times m \times n$. According to pilot experiments, a basic LPQ with $M = 7$ and an image grid size of 2×4 produces the best performance in our application.

As a family of LBP-based detectors, LBP and LPQ share some similar characteristics. For instance, both of them are static local appearance descriptors. To be more specific, they both compute the features on a very small neighbourhood surrounding pixels, and a binary number is attained. Finally a histogram is used to record the distribution of the occurrences of the possible binary patterns.

III. DYNAMIC LOCAL APPEARANCE DESCRIPTORS

Facial expressions are inherently dynamic processes. It has been shown that human interpretation of facial expressions depends heavily on the availability of the dynamics [3]. These cannot be described by the static appearance descriptors described in Section II. Instead, we can use the temporal extension of LBP, called LBP-TOP for this. Because LPQ outperforms LBP, both for face recognition [1] and in our own experiments (see Section V), we propose to extend LPQ in a similar fashion, and name it LPQ-TOP. Both descriptors result in a sparse encoding of appearance in small space-time video volumes.

A. LBP-TOP

Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) was proposed by Zhao and Pietikainen [23]. Originally, the textures were modelled with volume local binary patterns (VLBP), which are an extension of the LBP operator widely used in ordinary texture analysis, combining motion and appearance. Refering to the definition of LBP,

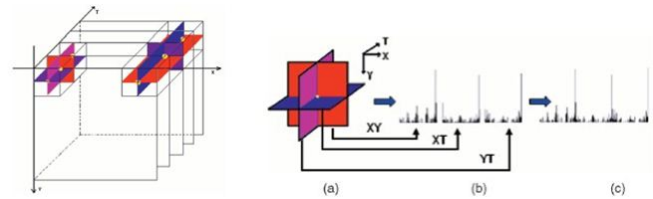

Fig. 3. Left: Three plane in spatio-temporal domain to extract neighbouring points. Right: Concatenated histogram from three planes [23]

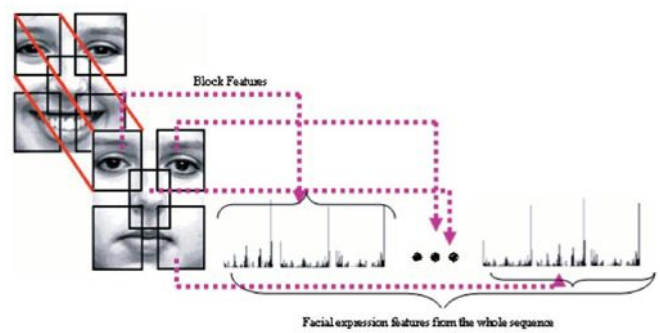

Fig. 4. The concatenated feature vector that extracted from each block to represent the whole sequence [23]

they define a Dynamic Texture (DT) V in a local neighbourhood of a monochrome DT sequence as the joint distribution v of the gray levels of $3P + 3(P > 1)$ image pixels, where P is the number of neighbourhoods.

In the proposed VLBP, a P neighbourhoods VLBP operator will result in a 2^{3P+2} dimension feature vector. When the number of neighbouring points increases, the number of patterns for basic VLBP will become very large. To make the approach computationally simple and easy to extend, a sparse space-time descriptor is proposed by computing LBP only on three orthogonal planes: XY, XT, and YT, and concatenating their results (shown in Fig.3).

The dynamic local appearance descriptor is denoted as LBP-TOP$_{P_{XY},P_{XT},P_{YT},R_{XY},R_{XT},R_{YT}}$, where $P_{XY}, P_{XT}, P_{YT}, R_{XY}, R_{XT}, R_{YT}$ are the number of neighbouring points in the XY, XT and YT planes and the radii in axes X, Y, and T respectively.

As we mentioned before, the LBP features computed over the whole face do not encode any spatial information. Following the idea of LBP and LPQ implementations, the image sequence is divided into several video volumes. The LBP-TOP histogram is extracted in each video volumes and concatenated into a single histogram (as shown in Fig.4). For example, A division into 5×10 video volumes will result in a feature vector of dimension $256 \times 3 \times 5 \times 10 = 8850$. A $LBP-TOP_{8,8,8,1,1,2}$ descriptor is adopted in this work.

B. LPQ-TOP

Following the idea of LBP-TOP, we extend LPQ to LPQ-TOP. The basic LPQ features are extracted independently from three orthogonal planes: XY, XT and YT, considering only the co-occurrence statistics in these three directions. To

be more specific, the phase information is computed locally in a window for every image position in three directions, and the phase of the four low-frequency coefficients are decorrelated and uniformly quantised in an eight-dimensional space. The histogram is obtained by accumulating the occurrence of quantised phase code in each direction, denoted as XY-LPQ, XT-LPQ and YT-LPQ. The histograms from the three orthogonal planes are concatenated to form a single histogram. Just as with LBP-TOP, the XY plane provides the spatial domain information while the XT and YT planes provide temporal information. Thus, by using this dynamic texture descriptor, both appearance and motion in three directions are considered. By using this method, the number of bins becomes $256 \times 3 = 753$ per space-time volume.

We recall that for the computation of LPQ, the phase information is extracted by using the 2-D short-term Fourier transform (STFT) computed over a rectangular neighbourhood N_x at each pixel position. One of the most important parameters in LPQ is the window size. Similar to the LBP operator, a smaller window captures more detailed facial feature information. Yet, possibly more unimportant information could be involved such as effects produced by illumination, personal characteristics and so on. It is not reasonable to use the same size for the rectangular windows in each of the three planes. Take an eleven frame image sequence with a resolution of 200×200 as example, using a rectangular size of 5 by 5, in the XY plane, the appearance might still be kept; however, the motion along the time axis changes dramatically in the XT and YT planes. Therefore, we set different rectangular size N_x in Eq. 2 for different planes. In other words, for the XY plane, the STFT is computed over a W_x by W_y rectangular neighbourhood at each pixel position, a W_x by W_t rectangular neighbourhood for XT and a W_y by W_t rectangular neighbourhood YT plane. So the novel descriptor is denoted as LPQ-TOP$_{W_x,W_y,W_t}$.

Recall Eq. 2. Suppose the coordinates of the central pixel $\mathbf{x}$ are given by (x_c, y_c, t_c) and that N_x is a M by N window, then $\mathbf{y} = (x, y, t)$. Hence, we obtain

for XY-LPQ
$x \in \{x_{c-(W_x-1)/2}, \dots x_{c-1}, x_{c+1}, \dots x_{c+(W_x-1)/2}\}$,
$y \in \{y_{c-(W_y-1)/2}, \dots y_{c-1}, y_{c+1}, \dots y_{c+(W_y-1)/2}\}$,
$t = t_c$;
for XT-LPQ
$x \in \{x_{c-(W_x-1)/2}, \dots x_{c-1}, x_{c+1}, \dots x_{c+(W_x-1)/2}\}$,
$y = y_c$,
$t \in \{t_{c-(W_t-1)/2}, \dots t_{c-1}, t_{c+1}, \dots t_{c+(W_t-1)/2}\}$;
for YT-TOP
$x = x_c$,
$y \in \{y_{c-(W_y-1)/2}, \dots y_{c-1}, y_{c+1}, \dots y_{c+(W_y-1)/2}\}$,
$t \in \{t_{c-(W_t-1)/2}, \dots t_{c-1}, t_{c+1}, \dots t_{c+(W_t-1)/2}\}$.

Now assume an $X \times Y \times T$ image sequence is given. Then $x_c \in \{0, ..., X-1\}, y_c \in \{0, ..., Y-1\}, t_c \in \{0, ..., T-1\}$. Note that some pixels in the boundary are omitted as their neighbourhood region is out of border. To calculate the LPQ-TOP distribution for this sequence, a histogram $H_{i,j}$ is

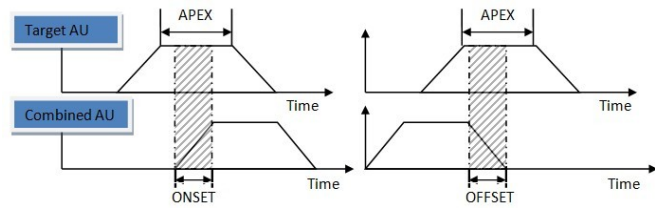

Fig. 5. The criterion of static data selection. The shaded areas are included in the dataset.

defined as

$$H_{i,j} = \sum_{x,y,t} I(f(x,y,t) = i),\ i = 0, 1, ..., 255, j = 0, 1, 2 \tag{5}$$

where $f_i(x, y, t)$ presents the LPQ code of the central pixel (x, y, t) in the jth plane which is computed by following the processes described in Eq. 2.

Also, because the computed dynamic texture features are of different spatial and temporal size, the histogram must be normalised to get a consistent description, so we write

$$N_{i,j} = \frac{H_{i,j}}{\sum_{k=0}^{255} H_{k,j}} \tag{6}$$

Hence, three 256 bin histograms, XY-LPQ, XT-LPQ and YT-LPQ are effectively obtained and concatenated to build the final LPQ-TOP feature vector. In our experiments, LPQ-TOP$_{7,7,3}$ is employed.

IV. DATA COLLECTION

In this work, the efficiency of the discussed descriptors is tested based on datasets collected from the MMI Facial Expression Database (MMI database) [18], and the SEMAINE database [7]. As LBP and LPQ deal with static images, while LBP-TOP and LPQ-TOP process image sequences, a separate dataset is collected for each.

The MMI database is a fully web-searchable collection of visual and audio-visual recordings of subjects displaying facial expressions which are FACS annotated. It includes 69 different subjects of both sexes (44 female), ranging in age from 19 to 62, having either a European, African, Asian, Caribbean or South American ethnic background. All fully FACS-coded recordings show facial expressions that are posed, and it is these data which will be used in this work.

To test our proposed AU detection methods on spontaneous facial expression data, we collect a dataset from the SEMAINE database. The SEMAINE database consists of a large number of emotionally coloured conversations. All the expressions of users are naturally induced by operators during the conversation. The dataset includes speech related mouth and face movements, and significant amounts of both in- and out-of-plane head rotations. All of these make the work even more challenging. Recently a small portion of the SEMAINE database has been FACS annotated by experts.

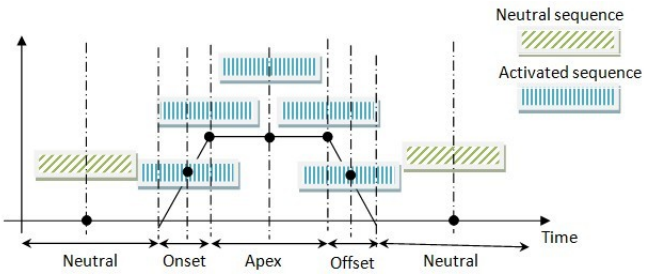

Fig. 6. The criterion of dynamic data selection. Each marked period results in one data element. Dynamic appearance descriptors are extracted from space-time volumes centered at salient moments indicated by the dots

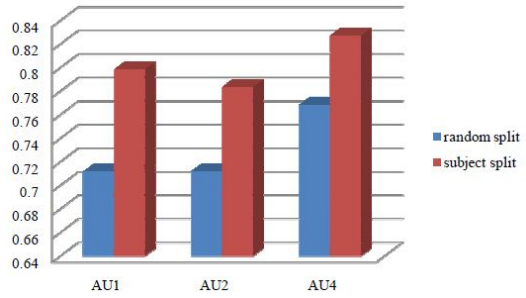

Fig. 7. Average F1-measures over all subjects based on different split approaches of parameter optimisation

A. STATIC DATASET

For the static dataset collection, we aim to collect a set of frames that is as sparse as possible, yet spans the appearance space of that AU completely. Note that when more than one AU is activated, facial actions can appear very different from when they occur in isolation. For example, AU1 and AU2 pull the brow up, whereas AU4 pulls the brows together and down using primarily the corrugators muscle at the bridge of the nose. The appearance of AU4 changes dramatically depending on whether it occurs alone or in combination with AU1 and AU2. In order to capture the appearance of each action unit as fully as possible and thus build a richer data space, we take in every video the first apex frames of each target AU, and all the apex frames where any other upper face AUs are in onset or offset (see Fig. 5). The shaded parts are the frames selected. However, AU combinations are not treated differently by the classifiers. In other words, each AU is recognised independently from all the others.

B. DYNAMIC DATASET

The dynamic dataset consists of a set of image sequences, which are extracted given a temporal window. Recall that the aim of AU detection is to know at any time whether an AU was present for that frame. This brings some difficult questions such as how long the temporal window we use should be and how to select the part of the video which best captures the appearance changes etc. One way of generating the training data is to first determine the window length, and then generate features for every frame. However, this may result in too many identical or at least very similar features.

Therefore, in our approach, we first define salient moments, namely the transition times between the different temporal segments and the midpoint of every AU phase. Secondly we generate an image sequence according to these key points and a temporal window. The key moments define where to apply our temporal window. As shown in Fig. 6, the vertical stripes rectangle shows activated image sequences and the diagonal stripes rectangle represents neutral image sequences in a video. Notice that the transition points between neutral and onset are omitted as the image sequences with half neutral and half apex frames may confuse whether to be classified as having an active AU or not. For testing a completely new video, we would generate the features for each frame. Here the optimal length of the temporal window is still an open question, and is of course a function at the frame rate. We will experiment with a window length of 0.445 seconds, i.e 11 frames here.

V. EVALUATION

We evaluated the two static and dynamic appearance descriptors on 442 videos taken from the MMI database, and 8 videos of approximately 5 minute conversations each from the SEMAINE database. As this is a user independent system for FACS Action Unit detection, the evaluation is done in a subject independent manner. Generalisation to new subjects is tested using leave-one-subject-out cross validation in which all images of the test subject were excluded from training. Hence no data from one subject appears in both the training and testing set.

A previous successful technique for facial expression classification is the Support Vector Machine (SVM) (e.g.,[2],[12]). In this work, we adopted SVM as classifiers for AU detection.

The evaluation is performed in three steps: feature normalisation, feature selection, and SVM classification. Firstly, the obtained histograms are normalised so that all the features lie in the range from -1 to 1. Next we use the GentleBoost algorithm to select the most informative features from the obtained histograms. So the problem space is reduced which in turn increases the classification accuracy [2]. Finally SVM classifiers are employed to perform AU activation detection.

When the GentleBoost algorithm is employed as a feature selector preceding a SVM classifier, at each stage a weak classifier is trained on a subset of the data consisting of a single feature and iteratively boosted to a strong classifier of higher accuracy. At each iteration, the weak classifier which minimizes the weighted error rate is selected, and the feature that this weak classifier represents is added to the list of selected features. An updated strong classifier is used to classify the training data, and the distribution is updated to increase the weights of the misclassified examples and reduce the importance of the others. This ensures that new features are selected that are contingent on the features that have already been selected, eliminating redundant information with each training round.

The SVM classification is achieved by the following steps. First, the instances with selected features are divided into two groups: a training set and a testing set. The SVM classifiers that we will use have a small number of parameters to set while training the classifier. As we want to find the optimal parameters for our problem in terms of classification performance, we employ a separate 5-fold cross validation loop each time we train a classifier in which we search for the optimal parameters.

Thus, while evaluating each fold, the training data is split into five subsets, four of which are used to train a classifier and one of which is used for testing. The partitioning could be done either randomly or in a subject-independent manner. An early-stage empirical study was carried out which compared the performance between random splits and subject splits. Fig. 7 illustrates the average F1 measures over all subjects obtained based on the different split approaches. It is obvious that subject independently splitting data during parameter optimisation generates much better SVM parameters. After optimisation, the SVM is trained using all training data and the optimal parameters. Separate binary classifiers, one for each AU, were trained to detect the presence of the AU regardless of co-occurring AUs.

Table I and Table II presents the AU recognition results using LBP and LPQ based on posed and spontaneous data. We can clearly see that LPQ outperforms LBP for most AUs. The importance is even clearer in the spontaneous data. The weighted average F1-measure from LPQ is 16% higer than that for LBP. As limited spontaneous data has been FACS-coded in the SEMAINE database, AU43 and AU46 are missing in Table II.

This difference between LBP and LPQ is even more prominent for the dynamic appearance descriptors. Table III shows the results obtained based on the LBP-TOP and LPQ-TOP descriptors. As we can see, in general, the system based on LPQ-TOP achieves a higher accuracy rate than LBP-TOP system. For AU45, LBP-TOP significantly outperforms LPQ-TOP. The reason for this is unknown, and this issue needs further investigation. Not taking AU45 in account, LPQ-TOP attained an average F1-measure of 82.9% and LBP-TOP an average of 75.8%, when averaging over the total number of positive examples.

Because different datasets were used for the static and dynamic appearance descriptors, we cannot directly use the results of tables I and III to make a comparison. Yet, since both the static and dynamic datasets are extracted from the same databases, we can compare all the tested descriptors by thresholding our prediction and assigning an unique label to a video. In other words, for each AU, if any frame or image sequence is classified as positive, that AU is said to be activated in the corresponding video. Note that this is done only for comparison purposes, and the results may not be optimal. Fig.8 presents the F1 measures over all videos for each AU as classified by the different methods. As we can see, the systems that utilise dynamic appearance descriptors outperform those that use static appearance descriptors. Our proposed spatio-temporal descriptor, LPQ-TOP, outperforms all other tested methods in terms of recognition accuracy for all AUs but AU45. Although it is hard to compare our results with those of others without controlled conditions, the results are a significant improvement when compared with [17], who tested on the same database.

TABLE I

AU DETECTION RESULTS USING LBP AND LPQ BASED ON POSED DATA TAKEN FROM THE MMI DATABASE

		LBP				LPQ			
AU	**n**	**CR**	**PR**	**RC**	**F1**	**CR**	**PR**	**RC**	**F1**
1	254	86.3	77.7	73.8	66.3	87.8	84.3	80.9	74.3
2	296	69.8	68.2	71.0	53.7	89.3	84.5	88.5	82.9
4	188	91.9	80.7	82.1	75.0	92.0	84.4	80.8	76.1
5	232	91.9	85.6	84.4	79.5	94.6	89.8	84.0	80.7
6	157	87.6	80.0	65.9	58.5	91.8	82.7	75.4	67.2
7	70	93.9	89.7	73.7	70.3	94.3	88.8	74.5	67.5
43	27	98.9	93.7	92.7	88.9	98.7	96.1	82.6	80.6
45	245	88.0	79.8	84.0	76.1	84.1	78.1	73.6	65.6
46	35	95.7	91.4	72.2	69.0	98.1	89.6	93.5	88.3
AVG	167	85.8	79.2	77.1	68.4	90.2	84.6	81.1	75.2

AU = Action Unit, CR = Classification Rate(%), PR = Precision(%), RC = Recall(%), F1 = F1-measure(%), n = number of positive examples, AVG = weighted average.

TABLE II

AU DETECTION RESULTS USING LBP AND LPQ BASED ON SPONTANEOUS DATA TAKEN FROM THE SEMAINE DATABASE

		LBP				LPQ			
AU	**n**	**CR**	**PR**	**RC**	**F1**	**CR**	**PR**	**RC**	**F1**
1	26	92.6	93.3	89.3	90.4	97.1	93.3	100	96.4
2	27	61.4	83.3	26.7	29.6	91.4	84.2	90.9	87.4
4	22	94.3	90.2	90.2	90.2	84.3	72.3	78.9	74.5
5	10	85.7	41.7	66.7	45.0	88.6	54.2	75.0	55.0
6	29	58.6	43.9	61.1	45.1	85.7	77.0	87.8	81.7
7	24	74.3	63.6	62.5	62.6	72.9	45.0	63.9	52.7
45	3	92.9	75.0	37.5	45.0	91.4	100	25.0	33.3
AVG	20	76.3	71.6	64.4	60.5	86.8	74.1	82.9	76.5

AU = Action Unit, CR = Classification Rate(%), PR = Precision(%), RC = Recall(%), F1 = F1-measure(%), n = number of positive examples, AVG = weighted average.

VI. CONCLUSIONS

We successfully implemented a robust and real-time AU detection system. We compared the static LBP and LPQ appearance descriptors with dynamic appearance descriptors LBP-TOP, and extend LPQ to LPQ-TOP. Results show that the systems based on LPQ generally achieve higher

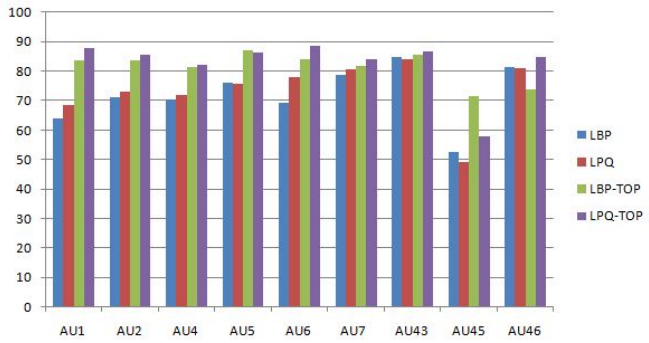

Fig. 8. The F1 measures (%) over all videos based on all tested descriptors

TABLE III

AU DETECTION RESULTS USING LBP-TOP AND LPQ-TOP BASED ON POSED DATA TAKEN FROM THE MMI DATABASE

		LBP-TOP				LPQ-TOP			
AU	**n**	**CR**	**PR**	**RC**	**F1**	**CR**	**PR**	**RC**	**F1**
1	148	91.0	83.5	83.1	77.5	90.5	86.9	90.0	85.6
2	138	88.8	81.1	77.7	73.6	87.8	84.5	85.1	79.4
4	129	89.2	82.8	75.6	72.2	88.3	85.7	84.4	81.2
5	50	92.3	89.0	85.7	79.5	94.0	89.6	87.1	83.2
6	58	95.6	85.4	86.4	82.6	97.2	88.9	93.4	87.2
7	48	90.7	93.4	73.1	73.8	94.1	89.8	80.2	80.9
43	61	96.6	85.0	87.9	79.2	97.7	89.1	92.3	86.3
45	664	86.4	83.1	78.6	75.3	76.1	63.4	63.2	48.1
46	98	92.7	89.5	77.0	73.6	92.8	89.9	88.4	82.4
AVG	103	89.0	84.1	79.4	75.5	84.2	75.9	75.9	66.3
AVG*	91	91.4	85.0	80.1	75.8	91.5	87.4	87.5	82.9

AU = Action Unit, CR = Classification Rate(%), PR = Precision(%), RC = Recall(%), F1 = F1-measure(%), n = number of positive examples, AVG = weighted average, AVG* = weighted average without considering AU45.

accuracy rate than LBP system, and that the systems that utilise dynamic appearance descriptors outperform those that use static appearance descriptors. Although the family of LPQ descriptors are more computationally expensive than the LBPs, they attain a higher recognition performance. All in all, the experimental results clearly show that our proposed spatio-temporal descriptor, LPQ-TOP, outperforms all other tested methods for the problem of FACS Action Unit analysis. Note that although we only applied the method to upper face AUs, the method can be readily used for all other AUs. The LBP-based version is freely available as part of the SEMAINE framework.

VII. ACKNOWLEDGMENTS

This work has been funded in part by the European Communitys 7th Framework Programme [FP7/20072013] under the grant agreement no 231287 (SSPNet). The work of Michel Valstar was also funded in part by the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE). The work of Maja Pantic is further funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

REFERENCES

[1] T. Ahonen, E. Rahtu, and J. Heikkila. Recognition of blurred faces using local phase quantization. In *Proc., Int'l Conf. on Pattern Recognition*, pages 8–11, 2008.

[2] M. Bartlett, G. Littlewort-Ford, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behaviour. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 223–230, 2006.

[3] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:1–12, March 2004.

[4] P. Ekman and W. Friesen. *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978. Palo Alto, California, USA.

[5] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32:1940–1954, 2010.

[6] R. Mattivi and L. Shao. Human action recognition using lbp-top as sparse spatio-temporal feature descriptor. In *CAIP '09: Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns*, pages 740–747, 2009.

[7] G. McKeown, M. Valstar, R. Cowie, and M. Pantic. The semaine corpus of emotionally coloured character interactions. pages 1079–1084, July 2010.

[8] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distribution. *Pattern Recognition*, 29(1):51–59, 1996.

[9] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution grey-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[10] V. Ojansivu and J. Heikkila. Blur insensitive texture classification using local phase quantization. In *In Proc. Int. Conf. on Image and Signal Processing*, volume 5099, pages 236–243, 2008.

[11] V. Ojansivu, E. Rahtu, and J. Heikkila. Rotation invariant local phase quantisation for blur insensitive texture analysis. In *Proc., Int'l Conf. on Pattern Recognition*, pages 8–11, 2008.

[12] C. Shan, S. Gong, and P. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2008.

[13] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2), 2001.

[14] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007.

[15] G. Tzimiropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki. Robust fft-based scale-invariant image registration with image gradients. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(10):1899–1906, 2010.

[16] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, page 149, 2006.

[17] M. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *ICCV-HCI'07*, pages 118–127, 2007.

[18] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. Int'l Conf. Language Resources and Evaluation, W'shop on EMOTION*, pages 65–70, 2010.

[19] M. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 3, pages 76–84, 2005.

[20] P. Viola and M. Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2010.

[21] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[22] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor wavelets-based facial expression recognition using multi-layer perceptron. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 454–459, 1998.

[23] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary pattern with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2(6):915–928, 2007.