A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modelling

Bihan Jiang, *Student Member, IEEE,* Michel Valstar, *Member, IEEE,* Brais Martinez, *Member, IEEE,* and Maja Pantic, Fellow, IEEE

Abstract—Both the configuration and the dynamics of facial expressions are crucial for the interpretation of human facial behaviour. Yet to date, the vast majority of reported efforts in the field either do not take the dynamics of facial expressions into account or focus only on prototypic facial expressions of six basic emotions. Facial dynamics can be explicitly analysed by detecting the constituent temporal segments of Facial Action Coding System's (FACS) Action Units (AUs) - onset, apex, and offset. In this work, we present a novel approach to explicit analysis of temporal dynamics of facial actions using the dynamic appearance descriptor Local Phase Quantisation from Three Orthogonal Planes (LPQ-TOP). Temporal segments are detected by combining a discriminative classifier for detecting the temporal segments on a frame-by-frame basis with Markov Models that enforce temporal consistency over the whole episode. The system is evaluated in detail over the MMI facial expression database, the UNBC-McMaster pain database, the SAL database and the GEMEP-FERA dataset in database-dependent experiments, and in cross-database experiments using the Cohn-Kanade and the SEMAINE databases. The comparison with other state-of-the-art methods shows that the proposed LPQ-TOP method outperforms other approaches for the problem of AU temporal segment detection, and that overall AU activation detection benefits from dynamic appearance information.

Index Terms—Facial dynamics, action unit detection, dynamic appearance descriptors, LPQ-TOP, temporal segment detection.

1 Introduction

Faces hold valuable clues to emotions and intentions of a person. Facial expressions are some of the most direct, naturally preeminent means for human beings to regulate interactions with each other [13]. They communicate emotions, clarify and stress what is being said, and signal comprehension, disagreement and intentions. Machine understanding of facial expressions could revolutionise user interfaces for artefacts such as robots, mobile devices, cars, and conversational agents [7], [32], and has therefore become a hot issue in Computer Vision and Pattern Recognition communities.

There are two main approaches to facial expression measurement in the field of psychology: message and sign judgement [7]. Message judgement aims to infer what underlies a displayed facial expression, such as affect or personality, while the aim of sign judgement is to describe the surface of the displayed behaviour, such as facial movement or facial component shape [32].

The Facial Action Coding System (FACS) [12] is the best known and most commonly used sign judgement approach developed for human observers to describe facial actions. It defines 32 atomic facial muscle actions named Action Units (AUs). With FACS, every possible facial expression can be described as a combination of AUs and a small set of Action Descriptors (as shown in Fig. 1). The latter causes appearance changes in the face but cannot be attributed to particular facial muscles. It would be possible to apply message judgement interpretation on the description of a facial expression in terms of AUs. For example, expressions typically associated with happiness contain AU6 and AU12 while those associated with sadness contain AU1, AU4 and AU15.

Fig. 1. Examples of upper and lower face AUs defined in FACS

Besides the configuration of facial expressions, their dynamics play an important role in the interpretation of human facial behaviour. For example, psychologists have found a difference in duration and smoothness between spontaneous and deliberate expressions, e.g. between polite and amused smiles [8], [11]. This has been confirmed in studies on automatic facial expression anal-

- *Bihan Jiang, Brais Martinez and Maja Pantic are with the IBUG group, Department of Computing, Imperial College London, UK. E-mail: bi.jiang09, b.martinez, m.pantic@imperial.ac.uk*
- *Michel Valstar is with the Mixed Reality Lab, School of Computer Science, University of Nottingham, UK. E-mail: michel.valstar@nottingham.ac.uk*
- *Maja Pantic is also with the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands*

ysis [17], [44], [49]. Facial expression dynamics are also essential for categorisation of complex mental states such as various types of pain and mood [2], [51]. Researchers have only started to learn the potential discriminative power of these dynamics.

One aspect of facial dynamics can be explicitly analysed by detecting their constituent temporal segments (i.e., onset, apex and offset) or intensity levels (i.e., 0 to 5). Recently there were several works reporting on automatic coding of facial dynamics and intensity of prototypic expressions of emotions (e.g., [9], [16], [19], [21], [24], [25], [37], [38], [40]). Given the agnostic nature of AUs, detecting the temporal segments/intensity levels of AUs is a more general and challenging problem. Yet, only few works have reported targeting this problem. We provide a detailed overview of these related works in Section 2.

In general, a facial expression recognition system consists of three steps: pre-processing, feature extraction and classification. The pre-processing step usually relates to face localisation, tracking and registration to remove the variability due to changes in head pose and illumination. Feature extraction is a crucial step for successful facial expression recognition. Consequently this has become a major focus of research in the field. The traditional goal of feature extraction is to convert pixel data into a higher-level representation of shape, motion, colour and texture, which minimises within class variations whilst maximising between class variations. Additionally this higher-level representation often grants some robustness to environmental conditions such as illumination or colour sensitivity variation in cameras.

Two traditional approaches for face image representation are geometry-based methods and appearance-based methods. Geometry-based methods employ the geometric properties of a face such as the positions of facial fiducial points, the distances between pairs of facial points or the velocities of particular facial points (e.g., [34], [48]). Appearance features aim to capture changes in face texture such as those created by wrinkles and bulges, as well as changes caused by facial motion (e.g., [4], [18], [54]). Typical examples include Gabor filters, Haar-like filters, and other image filters. Appearance feature extraction methods can be applied to the whole face, specific face regions, or local patches around some fiducial points. The latter two approaches effectively combine geometric-based and appearance-based methods and are referred to as hybrid methods.

Facial expression recognition is by definition about recognising changes in the face, i.e., it is essentially facial action detection. To date, most appearance-based facial expression recognition systems use only static appearance descriptors, which means the appearance changes and their associated temporal information are completely ignored. This paper addresses this limitation of the state of the art by extending the static appearance descriptor Local Phase Quantisation (LPQ) [31] to temporal Three Orthogonal Planes (TOP), inspired by a similar extension of Local Binary Patterns, LBP-TOP [54]. The resulting spatio-temporal appearance descriptor LPQ-TOP is applied to detect the latent temporal information representing facial appearance changes and explicitly model facial dynamics of AUs in terms of their temporal segments. In this work we further show that spatio-temporal appearance descriptors are suitable for capturing the temporal segments of AUs, and they add valuable information with respect to static appearance descriptors.

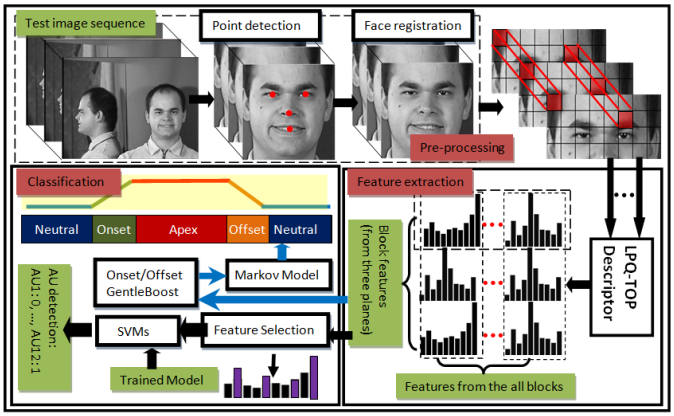

Fig. 2. Outline of the proposed fully automated system for recognition of AUs and their temporal activation models.

Fig. 2 shows an overview of our proposed system. Faces are detected and registered using the automatic point detector presented in [29]. The registered frames are used to extract dynamic appearance features in a holistic manner from the full face region. More specifically, a block-based representation is used. The resulting features are analysed by a set of classifiers, either trained to detect the activation of a target AU, or to detect its temporal segments. When dealing with temporal segments, a Markov Model (MM) is applied to impose temporal consistency on the assigned labels.

In this work we also investigate various approaches to select data to learn from. Specifically, creating a representative training data set to learn from is by no means a trivial task when it comes to AU detection and AU temporal segmentation. In this work we present a novel heuristic method and compare its efficiency with two standard methods of training-data selection: random sampling and bootstrapping. Our proposed heuristic training-data-selection approach gives comparable results to other approaches at a low computational cost.

To summarise, the contributions of this paper are as follows:

- We propose the use of spatio-temporal features for AU analysis. We show that this outperforms both the use of spatial only and temporal only (i.e. motion-based) approaches. This is consistent over multiple databases and for two appearance descriptors;
- We show that dynamic appearance descriptors en-

able the detection of AU temporal segments. Until now, only motion-based and geometric-based features could be used successfully for this task;
- We propose a heuristic training-data selection approach and demonstrate that it outperforms other standard data-selection techniques.

The remainder of this paper is organised as follows. Section 2 provides an overview of the related research on facial dynamics analysis. Section 3 presents a detailed explanation of the methodologies used in our approach. Section 4 describes the utilised data sets, the evaluation setup and parameter optimisation. Section 5 discusses the evaluation results. Section 6 concludes the paper.

2 Related work

In this section, we explore the existing approaches for automatic analysis of AUs and their dynamics.

2.1 Automatic Action Unit detection

Historically, the first attempts to automatically detect AUs in face images were reported by Bartlett et al. [5], Lien et al. [23] and Pantic et al. [35]. The focus in the field was first on automatic recognition of AUs in static images and then image sequences picturing facial expressions produced on command. A number of promising prototype systems that can recognise AUs in either (near-) frontal view face videos or profile face videos were reported (e.g., [42], [4] and [46]). This focus has shifted to automatic AU detection in spontaneous facial expressions [15], [52]. A number of methods have been proposed for AU detection in facial displays of pain [19], [24], [25]. For example, [25] uses geometric, appearance and hybrid features from Action Appearance Model (AAM) to detect AUs and pain at a frame-by-frame level; [19] combined different shape (facial landmarks), appearance (DCT and LBP) features and their fusion with regression techniques for the same problem.

Recently the first facial expression recognition challenge was organised, to take stock of how far the AU detection field has come. The challenge results indicated that the field still has some way to come before AU detection can be performed effectively on real world data in a robust manner [45].

2.2 Automatic Action Unit dynamics analysis

To the best of our knowledge, only five studies reported on automatic detection of AU temporal segments in frontal- [48], [22], [33], [36] and profile-view [34] face videos. The works presented in [33], [34] employ rule-based reasoning and geometric based features to encode AUs and their temporal segments. In [22], the authors use motion-based features and combine GentleBoost classifiers and Markov Models. More specifically, the authors use a nonrigid dense registration based on Free-Form Deformations to capture the facial motion between frames. [48] combined Support Vector Machine and Markov Models with geometric-based features. The authors detect and track a set of 20 facial points, of which the relative positions and displacements are calculated and used as features. As in [22], Markov Models are used to enforce temporal consistency on the assigned labels throughout the sequence. Recently, [36] used appearance features and their proposed model the Laplacian-regularised Kernel Conditional Ordinal Random Field model (Lap-KCORF) for the recognition of AUs' temporal segments. This model takes into account ordinal relations between the segments.

There is a small number of works that focus on the intensity estimation of AUs. Bartlett et al. [4] proposed to measure the AU intensity in posed and spontaneous facial expressions. They use distances to the SVM separating hyperplane as a direct measure of the intensity levels of AUs. Mahoor et al. [27] treat this problem as a multi-label classification problem. Hence they trained six one-vs-all SVM to classify a specific frame into one of the six FACS intensity levels. Savran et al. [3] proposed a novel intensity estimation scheme using 2D and 3D images. The scheme is based on regression of selected image features. In [19], the authors proposed a three-step approach to continuous pain intensity estimation based on Relevance Vector Regression.

There are other works that use aspects of the temporal dynamics of facial expression such as the speed of a facial point displacement or the persistence of facial parameters over time. However, this was mainly done either in order to increase the performance of facial expression analysis (e.g., [43], [53], [18]) or in order to report on the intensity of the shown facial expression (e.g., [53]). To date, the work by Tong et al. [43] is the only one that models the semantic and temporal relationships between AUs forming a facial expression, although it does not explicitly exploit AU facial dynamics in terms of temporal segments.

3 Methodology

This chapter presents the details of our approach, an outline of which is shown in Fig. 2. The implementation of the method described in this work, to wit the LPQ-TOP-based AU detector, is freely available and can be downloaded from http://ibug.doc.ic.ac.uk/resources/temporal-based-action-unit-detection/.

3.1 Preprocessing

In order to locate the face in an input frame and remove unwanted transformations such as rotation and translation, a version of the point detector described in [50] is adopted. A Procrustes transformation (i.e. a combination of translation, rotation and isotropic scaling) is computed by aligning the coordinates of the left eye, right eye, nose and mouth to a set of anchor points. The anchor points correspond to the location of the same facial components in a prototypical frontal face. The input

Fig. 3. Preprocessing illustration

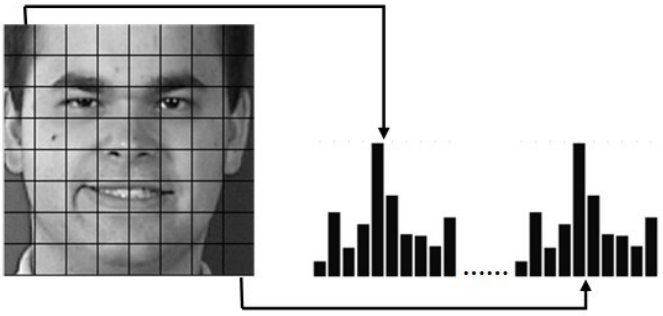

Fig. 4. The concatenated feature vector that extracted from each face block

frame is subsequently cropped to be 200×200 pixels (see Fig. 3). This step eliminates in-plane head rotation and addresses individual differences in face shapes. No effort was made to address illumination issues.

3.2 Appearance-based feature extraction

3.2.1 Local Phase Quantisation

The Local Phase Quantisation (LPQ) operator is a texture descriptor robust to image blurring [31]. 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

$$F(\mathbf{u}, \mathbf{x}) = \sum_{\mathbf{y} \in N_x} f(\mathbf{x}\text{-}\mathbf{y}) e^{-j2\pi \mathbf{u}^T \mathbf{y}} = \mathbf{w}_\mathbf{u}^T \mathbf{f}_\mathbf{x} \qquad (1)$$

where $\mathbf{w_u}$ is the basis vector of the 2-D DFT at frequency $\mathbf{u}$, and $\mathbf{f_x}$ 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 ($a = 1/M$ 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

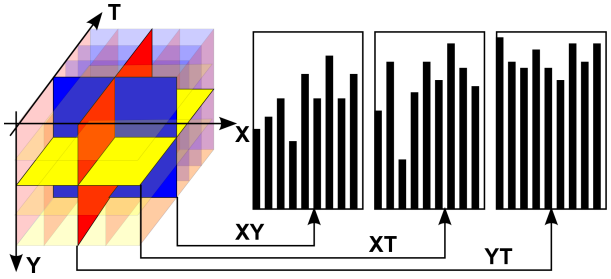

Fig. 5. Concatenated histogram from three planes

of each component in F_x. This is done by using a simple scalar quantiser

$$q_j = \begin{cases} 1 & \text{if } g_j \geq 0 \text{ is true} \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

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

$$f_{\text{LPQ}}(x) = \sum_{j=1}^{8} q_j 2^{j-1}. \qquad (3)$$

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

Histograms discard all information regarding the spatial arrangement of the patterns. In order to preserve some of this information, we divide the face region into m local regions, from which LPQ histograms are extracted and then concatenated into a single feature histogram (see Fig. 4). Originally proposed in [1], block-based representations have been adopted in most existing studies using holistic histogram-based face representation (e.g., [18], [39]). An image divided into $m \times n$ blocks will produce a feature vector with dimension of $256 \times m \times n$.

3.2.2 LPQ-TOP

To extend the LPQ descriptor to the temporal domain, the basic LPQ features are extracted independently from three sets of orthogonal planes: XY, XT and YT, considering only the co-occurrence statistics in these three directions, and stacking them into a single histogram (see Fig. 5) [54]. The XY planes provide the spatial domain information while the XT and YT planes provide temporal information. This method results in $256 \times 3 = 753$ bins per space-time volume. Note that features are extracted from all possible XY, XT, and YT planes, not just the three central planes depicted in Fig. 5. A similar approach has been adopted to extend LBP descriptor to the temporal domain [54], and this motivated our current work (see also [18]).

One important parameters for the LPQ descriptor is the neighbourhood size N_x. It is not reasonable to use the same rectangular neighbourhoods size of the spatial

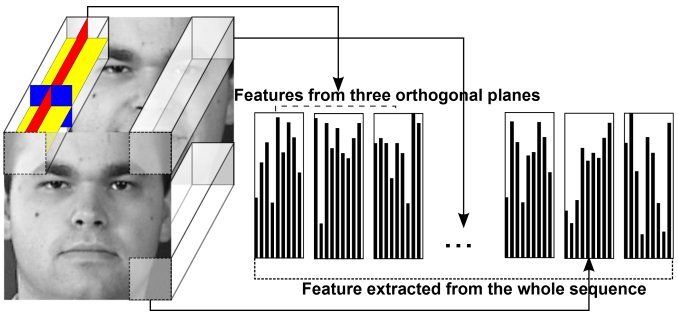

Fig. 6. The concatenated feature vector that extracted from each block to represent the whole sequence

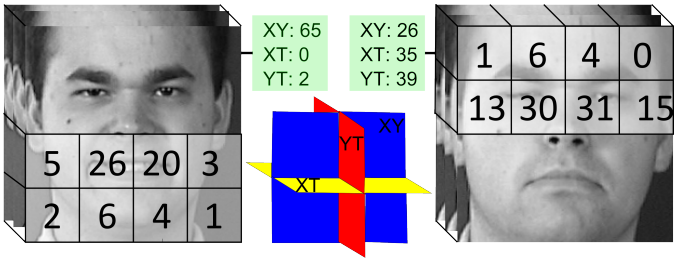

Fig. 7. Spatial localisation of the selected features. Each square is a face sub-region, and the number indicates how many features are taken from it. The number of features selected for each canonical plane is also shown. AU12 (smile) is depicted on the left, AU45 (blink) on the right. The MMI database was used for training.

plane and the two temporal planes. For example, we work with a face resolution of 200×200 pixels, while the frame rate is 25 fps. As these magnitudes are not directly comparable, we set a different rectangular size N_x for the different planes, denoted by W_x, W_y and W_z. That is to say, the XY descriptor is computed using a W_x by W_y rectangular neighbourhood at each pixel position. Finally, the histograms from each plane are normalised independently so that each adds to 1.

3.3 Classification

In this paper, two problems are addressed: AU activation detection and AU temporal segment detection. Note that AU activation detection is a binary classification problem, with highly unbalanced data. Temporal segment detection, on the other hand, is a multi-class problem. In this section, we will present the methodologies we use to deal with these two problems.

3.3.1 Action Unit Activation Detection

For this problem, a SVM is used as the binary classifier, while a GentleBoost algorithm is adopted preceding the SVM training for selecting the most relevant features. That is to say, we first train a GentleBoost classifier for each AU, and keep the subset of features used by it. Then, the SVM classifier for each AU is trained only on the feature subset selected by the GentleBoost classifier. This gives a reasonable balance between speed and complexity [22].

The SVM requires the optimisation of the error-insensitive margin (typically denoted by ϵ), the slack variable (typically denoted by C), and potentially any hyperparameters controlling the kernel function. This is done through a grid search strategy, where a separate subject-independent crossvalidation loop is performed within the main evaluation cross-validation loop to obtain the performance for each set of parameters.

Regarding the kernels tested, we compare the most commonly used ones in the literature (linear, plynomial, and rbf), plus the recently proposed intersection kernel [28], which is designed for histogram-based features. As our dynamic appearance descriptor LPQ-TOP is histogram-based, we expect this kernel to be very suitable for our problem.

More specifically, the intersection kernels is defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \sum_k min(\mathbf{X_i}(k), \mathbf{X_j}(k)) \quad (4)$$

Results on the performance of each kernel are presented in Section 5.

3.3.2 Action Unit Temporal Segmentation

For temporal segment detection, a dedicated one-versus-all GentleBoost classifier is trained for each AU and each temporal segment characterised by motion (i.e., onset and offset), as we experimentally found it to perform better than SVM for this task. As an example, Fig. 7 illustrates the distribution of the features used by GentleBoost for AU12 (smile) and AU45 (blink). As expected, most features are selected around relevant regions. It can also be seen that for AU12, 97% of the features are selected from the spatial domain (XY) while for AU45 74% of the features are from the temporal domain (XT and YT). This is in agreement with the finding that some AUs can be detected using static features only and for others dynamic features are crucial; for example, the only difference between AU43 (eye closure) and AU45 (blink) lies in the temporal domain (i.e., the duration of eye closure).

We combine the GentleBoost classifiers with a Markov Model (MM) in order to obtain temporal consistency over the assigned labels. MM is one of the most popular approaches to model time for classification problems [6]. In particular, a hidden node's variable can take one out of four values, each corresponding to a temporal phase. The relation between hidden nodes is modelled as a first order Markov chain and, therefore, its state depends on the state of the previous node (through the transition probability) and on an observation node (through the observation probability) only. Following [22], as discriminative classifier we use GentleBoost to model the relation between the hidden labels and the observation nodes instead of the traditional generative

modelling. This has been proved effective in practice, as discriminative models are usually able to learn class boundaries effectively with less training data and with feature vectors of higher dimensionality. Finally, the most likely sequence of hidden labels is found using the standard Viterbi algorithm.

4 EXPERIMENTAL SETUP

4.1 Facial expression data sets

The **MMI database** [50] and the **Cohn-Kanade (CK) database** [20] contain audio-visual recordings of subjects displaying posed facial expressions in frontal or near-frontal head poses and under controlled lighting conditions. They provide recordings of a wide range of AUs, as they are displayed on demand. The MMI database contains 264 videos of 10 subjects fully FACS-coded in terms of AU activation and temporal segments by two FACS experts. For the CK database, the videos are very short. They are typically 18 frames long, which means that the optimal window length of dynamic descriptors might be too large to be applied. Among the 187 available sequences, there are only 55 sequences longer than the minimum required of 21 frames. In addition the offset segment is not included in this database. Therefore, we only use this database for cross-database evaluation of AU activation detection.

One of the main criticisms of most existing works on automatic facial expression recognition is that they are based on deliberate and often exaggerated facial expressions which rarely occur in real life [32], [52]. Spontaneous facial expressions differ from posed expressions both with respect to which muscles are activated and in the dynamics of the muscle activations [2]. For this reason we also evaluate the proposed methodology on spontaneous expressions.

The **UNBC-McMaster Shoulder Pain Expression Archive Database** (UNBC-McMaster pain database) [26] is a publicly available database designed for the purpose of pain analysis. It contains a total of 129 participants (63 male) with shoulder pain. Although the camera is placed in front of the subjects, changes in head pose are common. Only the AUs that have been implicated as possibly related to pain were FACS-annotated (i.e. AU4, AU6, AU7, AU9, AU10, AU12, AU20, AU25 and AU26). The authors also provide the locations of a set of facial points per frame, obtained with a facial point tracker.

The **SAL dataset** [10] and the **SEMAINE datasets** [30] contain displays of spontaneous expressions recorded in a natural environment. The expressions were elicited in human-computer and human-human conversations, respectively. The head poses are mostly frontal or near-frontal due to the nature of the interaction, although the subjects can move freely. In SAL, ten subjects were recorded. After removing the speech sections, 77 sequences are left to perform our evaluation studies on. For four subjects, the data have been FACS-coded on a frame-by-frame basis in terms of temporal segments. For the other six subjects only AU activation coding exists. For the SEMAINE database, sparse annotations of AU activation is available for 10 sequences. This means that for a few frames per video AU annotations are available, with the frames that are annotated usually far apart. We use these 10 sequences for the purpose of cross-database AU activation detection evaluation.

The **GEMEP-FERA challenge dataset** is a subset of the GEMEP database. It contains 158 sequences of 10 subjects, who are professional actors using the Russian method of acting, i.e. evoking the feeling they are supposed to portray from experience, and then act on that feeling. It contains significant non-frontal head poses, and it is a very challenging dataset. It is split into a training set and a testing set. Frame-based AU activation has been annotated, but only the labels of the training set are publicly available. As there is no temporal segment annotation available for this data, we will only use it for the evaluation of AU activation detection.

All databases are recorded indoors, and have controlled even lighting. In addition to the existing office strip lights, the SAL and SEMAINE databases used indirect frontal illumination to reduce the effect of cast-shadows created by overhead lighting.

4.2 Training-Data Selection

The selection of the optimal set of training data is an important aspect when using machine learning techniques. The goal of training-data selection is to collect a set of examples that is as sparse as possible, yet spans the problem space completely.

The simplest way to deal with this problem is to randomly select the positive and negative examples with uniform probability. This however can result in including positive and negative instances whose feature values differ very little from each other (e.g., frames near the beginning of onsets and end of offsets), as well as inclusion of highly redundant examples, and overrepresentation of the most common patterns. Bootstrapping is a common and more sophisticated approach [41]. By iteratively adding the mis-classified examples to the training set, it refines the optimum hyperplane between positive and negative examples. However, this process can be extremely computational intensive. We propose a heuristic approach to the training-data selection, and compare it with these two standard data-selection techniques: uniform random sampling and bootstrapping.

Uniform Random Sampling - In this method, every instance has the same probability to be selected. We uniformly select n training instances from the original training set. In this formulation, n is the mean of the number of instances selected by heuristic approach and bootstrapping strategy.

Bootstrapping - Let us split the training database into three different datasets without mixing subjects: A, B and C. The training-bootstrapping algorithm used in this work then follows as:

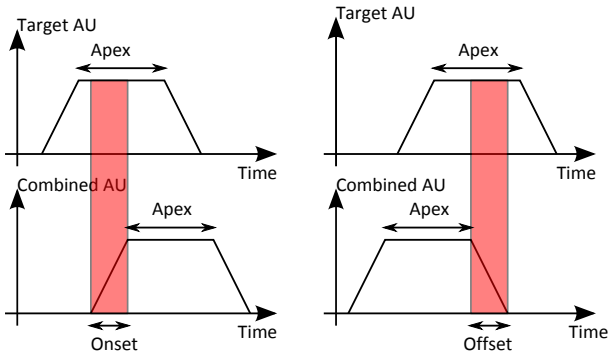

Fig. 8. The criterion of static data selection. The shaded areas are included in the dataset.

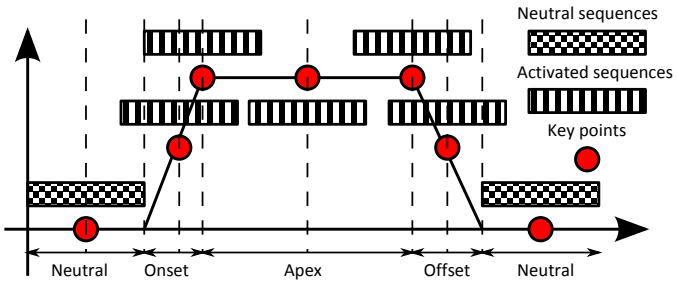

Fig. 9. The criterion of dynamic data selection. Each marked period results in one data element. Dynamic appearance descriptors are extracted from space-time volumes centred at salient moments indicated by the dots

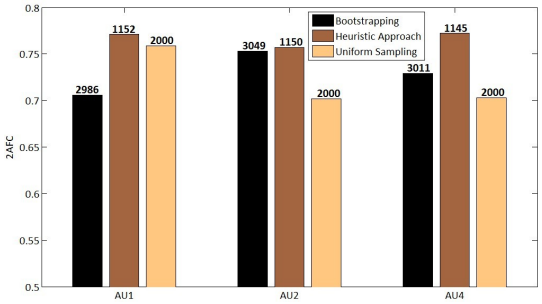

Fig. 10. Average 2AFC scores based on different training-data selection approaches tested on the MMI database. The number above each bar indicates the number of selected instances

1) Randomly select k samples from dataset A and B to get a and b respectively.
2) Train on a to get classifier c_a and test on B.
3) Gather the mis-classified examples of B and add them to b.
4) Train on b to get classifier c_b and test on A.
5) Gather the mis-classified examples of A and add them to a.
6) Test on C using classifier c_a and c_b.
7) Use the mean of the predictions on C i.e., decision values obtained from classifier c_a and c_b to get the performance f.
8) Terminate if f does not increase for more than three iterations to avoid local maximum. Otherwise, go to step 2.
9) Concatenate set a and b.

Note that the choice of k largely depends on the data. In this case, we start from the apex frames in the dataset. In each iteration, the false negatives and 1% of the false positives are added. We only add a subset of false positives because the dataset we utilised is highly unbalanced.

Heuristic approach - The heuristic approach we propose distinguishes two different cases, and produces two datasets for AU detection. One is the static dataset which selects training data for static appearance descriptors, the other is the dynamic dataset for dynamic appearance descriptors. For the static dataset, we note that when more than one AU is activated, facial actions can have a very different appearance than when they occur in isolation. These AUs are known as non-additive AUs. In order to capture the appearance of each action unit as fully as possible and thus build a richer data space, we take the first apex frames of each target AU, as well as other apex frames where any other upper face AUs are in onset or offset, as they produce the appearance changes (see Fig. 8, where the shaded parts are the frames selected). Through this approach, AU combinations can be expected to be detected by performing independent detections of the forming AUs.

The dynamic dataset consists of a set of spatio-temporal video segments extracted using a pre-defined temporal window. To avoid repetitive patterns and to reduce the potential number of examples, we first define salient moments as the transition times between the different temporal segments and the midpoint of every AU phase. Then a space-time volume centred at the salient moments is used to extract the appearance features. As shown in Fig. 9, the vertical striped rectangles show activated space-time volumes and the checkboard rectangles represent neutral space-time volumes 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 produce a pattern too similar to the negative class, and would confuse the classifiers.

Fig. 10 shows the average 2AFC scores based on different types of training selection approaches as discussed above. The appearance descriptor LPQ and intersection kernel have been used. It is clear that the heuristic approach achieves the highest accuracy among these three methods. The reasons for the poor performance of bootstrapping may lie in the highly unbalanced training data. The bootstrapping method iteratively trains a classifiers to find the optimal training set. This process is much more time-consuming than the heuristic approach. Moreover, note that the number of selected training instances directly affects the computational costs of the following steps, i.e., feature selection and classifier training. On average, 3000 instances are selected by bootstrapping,

1150 instances are selected by the heuristic approach, and for random sampling we randomly selected 2000 examples. Hence, the heuristic approach performs best, not only in terms of performance, but also in terms of computational complexity and memory consumption.

4.3 Evaluation setup

As our approach is intended to be a subject independent methodology for FACS Action Unit analysis, the evaluation is done in a subject independent manner. The ability of the system to generalise to new subjects is evaluated by using 10-fold subject-independent cross-validation when performing tests within the same dataset, and by training and testing on completely disparate datasets when performing cross-database experiments. The training instances are selected using the proposed heuristic approach described in Section 4.2 in all experiments.

The performance measure used in this work is the 2-alternative forced choice task (2AFC). The percentage of correctly classified examples in a 2AFC evaluation framework is equivalent to the area under the ROC curve (AUC) [14], and can be computed more efficiently than the AUC itself. Another performance measure typically used in the literature is the F1-measure, which is the harmonic mean of precision and recall. The F1-measure suffers from the problem of not crediting true negative detections. The 2AFC score does take the true negative rates into account and is therefore preferred. In a practical application, a vast majority of frames will not have the target AU active; therefore measuring the true negatives is very important. In this work we will only use the F1-measure to allow comparison with other works.

4.4 Parameter optimisation

Our approach requires the optimisation of the choice of SVM kernel and its parameters, the spatial block-size of the appearance descriptors (see Sec. 3.2.2), and the temporal window length of the appearance descriptors. It is important to note that the only extra parameter with respect to many other methods is the temporal window length, although often the optimisation process for some of these parameters is not mentioned explicitly.

To find the optimal height and width of the blocks in the histogram grid, we tested the average performance using different block sizes. We use the static descriptor as the block size relates to the spatial information only. We tested on grids of 4×4, 8×8 and 10×10 blocks (the more blocks an image is divided into, the fewer pixels each block contains). We found that 10×10 produces better results. We directly apply our results from the static descriptors to the dynamic descriptors.

The kernel parameters are optimised through a grid search, where a second subject-independent cross-validation loop is applied to the training set of every fold of the outer cross-validation definition. To select the best-performing kernel, we computed the average 2AFC score using 9 common AUs, while LPQ-TOP features were used. The results yielded similar results across kernels, although the intersection kernel performed slightly better than the rbf kernel. As the histogram kernel is also more computationally efficient than the RBF kernel, that's what we selected for our experiments.

As different AUs have different dynamics, we explored the optimal window length for each AU independently. The window lengths tested ranged between 3 and 21 frames. We found that the optimal window length further depends on the database setting, so we optimised this parameter during the training stage for each experiment, and include the optimal figures obtained in the corresponding results tables. Again, the optimal window length was determined independently of the test data.

The optimal window lengths used to detect temporal segments are found independently of those for AU detection, as the temporal segments are strictly shorter in duration than the full AU episodes. Different AUs also have different segment durations.

5 Evaluation results

We conducted three sets of experiments to evaluate the performance of our method. The first set of experiments is designed to evaluate its performance regarding frame-based AU detection. More specifically, we provide quantitative performance evaluations for posed and spontaneous databases, and we show that the use of dynamic features improves the results of AU detection when compared to their static counterparts. The MMI database is used to provide performance results for posed AUs, while the UNBC-McMaster pain database, the SAL database and the GEMEP-FERA dataset are the benchmarks for spontaneous expressions.

The second round of experiments is targeted at detecting the temporal AU segments. While the previous problem was binary (an AU is active or not), now the output of the algorithm can belong to 4 classes (neutral, onset, apex or offset). For this evaluation we use the MMI, UNBC-McMaster pain and SAL databases, which are used by other state-of-the-art methods as well. However, only the MMI database contains sufficient annotations in terms of temporal segments to perform experiments. It is therefore complicated to quantify performance of temporal segment detection in spontaneous settings. In order to do so, we follow the same procedure as in [22], where the output of the temporal segment detection was then converted into a binary output representing AU activation. That is to say, if either onset, apex or offset was detected, then the corresponding binarised output used for comparison would be 1.

The last set of experiments explores how generalisable the results are. To this end, we perform a series of cross-database experiments; we trained on the MMI database and tested on the Cohn-Kanade database, and we trained on SAL and tested on the SEMAINE database.

In this section, comparisons between LBP-TOP and LPQ-TOP are carried out as to show the superior performance on average of LPQ-TOP. Furthermore, we compare relative performance boost obtained by using dynamic features by comparing the performance of LBP and LBP-TOP and LPQ with LPQ-TOP respectively.

5.1 Frame-based Action Unit detection

TABLE 1
AU activation detection results (2AFC) using LBP, LPQ, LBP-TOP and LPQ-TOP based on posed data taken from the MMI database. n is the number of tested videos, θ_1 is the optimal window length for LBP-TOP and θ_2 is the optimal window length for LPQ-TOP.

AU	n	LBP	LPQ	θ_1	LBP-TOP	θ_2	LPQ-TOP
1	13	0.838	0.766	19	0.815	21	**0.850**
2	12	0.794	**0.832**	21	0.809	21	0.822
4	33	0.785	0.799	21	0.819	19	**0.828**
5	12	**0.837**	0.831	11	0.810	15	0.825
6	19	0.599	0.702	15	0.805	21	**0.810**
7	10	0.577	0.534	11	**0.686**	19	0.678
9	11	0.812	0.770	21	0.889	11	**0.959**
10	14	0.807	0.820	11	**0.892**	21	0.877
11	17	0.858	0.903	11	0.978	19	**0.983**
12	18	0.884	0.892	21	0.937	21	**0.958**
13	9	0.887	0.857	21	**0.986**	21	0.973
14	16	0.804	**0.882**	15	0.824	19	0.878
15	12	0.831	**0.862**	11	0.841	15	0.813
16	14	0.801	0.850	15	0.894	21	**0.952**
17	93	0.691	0.820	21	0.811	19	**0.828**
18	21	0.738	0.758	21	0.881	11	**0.904**
20	11	**0.852**	0.845	21	0.776	21	0.740
22	11	0.801	0.871	15	0.860	19	**0.886**
23	12	0.654	0.608	21	**0.785**	21	0.752
24	19	0.690	0.640	15	**0.774**	21	0.754
25	75	0.758	0.784	15	**0.805**	21	0.795
26	33	0.700	0.715	19	0.851	19	**0.885**
27	13	0.849	0.880	21	0.831	21	**0.947**
28	35	0.856	0.905	11	**0.908**	15	0.863
43	11	0.754	0.848	21	0.920	21	**0.968**
45	108	0.613	0.683	7	**0.896**	7	0.838
46L	22	0.887	0.896	11	0.901	7	**0.903**
46R	11	0.876	0.881	11	0.912	7	**0.941**
AVG		0.780	0.801		0.853		**0.865**

The 2AFC scores obtained with our method on the MMI database in terms of AU detection are shown in Table 1. It is possible to draw two conclusions from the obtained results. Firstly, the best-performing feature for this task is the LPQ-TOP, although this is not the case for all AUs. Secondly, using dynamic appearance descriptors provides a significant performance boost respect their static counterparts, with a similar boost obtained for both. The improvement in performance of dynamic features can be illustrated for the case of AU45 (blink). As can be seen from the table, LBP-TOP performance is 28.3% higher than that of LBP for AU45. The reason that the actual difference between AU43 (eye closed) and AU45 lies in the temporal domain.

We used the UNBC-McMaster pain database and the SAL database in order to evaluate the performance of our method for spontaneous expressions. The set of tested AUs is smaller, as the number of AUs occurring in spontaneous expressions for a specific scenario (in this case pain/dyadic interactions) is smaller than with acted AU displays. For the UNBC-McMaster database, the 10 AUs which have been implicated as possibly related to pain are tested.

The results using LPQ-TOP, as well as those using similarity normalised shape (SPTS), canonical normalised appearance (CAPP) and a hybrid method reported in [25] have been presented in Table 2. For a fair comparison, the same pre-processing step reported in [25] was used, with the LPQ-TOP features extracted from the normalised appearance. As we can see, the system using LPQ-TOP outperforms those using geometric (SPTS), static appearance (CAPP) and hybrid features on average, particularly for AU4 and AU43. The performance of dynamic appearance feature LPQ-TOP is 5.5% higher than that of the static appearance feature CAPP. Furthermore, geometric features were found to be particularly suitable for the detection of AU25 and AU26. Note that while the difference between LPQ-TOP and the Hybrid appearance/geometric approach of [25] is negligible, our system does not leverage geometric features in the Machine Learning phase. We expect enhancing the LPQ-TOP with geometric features would result in a similar performance increased observed between the SPTS, CAPP and Hybrid approaches.

For the SAL database, only 10 AUs are evaluated, being the only ones that occurred five or more times in the training data, and it is the same set as used in [22]. As can be seen from Table 3, the achieved 2AFC score is 0.81, while the average performance on the MMI database for this same subset of AUs is 0.83. Therefore, the performance loss with respect to posed expressions is marginal. This is despite the common understanding that spontaneous expressions are more challenging.

TABLE 2
Results (2AFC) for Testing the System on the UNBC-McMaster pain database for the LPQ-TOP in terms of frame-based AU detection accuracy. For comparison, results from [25] by using the similarity-normalised shape (SPTS), the canonical appearance (CAPP) and both have also been presented. θ is the optimal window length for the LPQ-TOP descriptor, N is the number of frames that contain an AU.

AU	θ	N	LPQ-TOP	SPTS	CAPP	Hybrid
4	15	1074	**91.6**	72.5	60.0	57.1
6	11	5557	80.6	80.1	85.1	**85.4**
7	20	3366	74.0	71.3	**82.6**	80.4
9	11	423	85.0	75.1	84.1	**85.3**
10	15	525	86.9	87.9	83.2	**89.2**
12	19	6887	79.8	79.4	84.6	**85.7**
20	15	706	68.9	75.7	61.7	**77.9**
25	19	2407	74.3	**78.8**	70.9	78
26	11	2093	61.8	**73.5**	54.7	71
43	10	2434	**95.7**	83.1	86.7	87.5
AVG	-	-	**79.9**	78.0	75.4	79.8

Further evaluation has been carried out on the GEMEP-FERA dataset. Unlike the SAL database, it contains a large number of displays in non-frontal head poses. This is particularly challenging for texture-based methods, and especially for holistic methods. This is

TABLE 3
Results for Testing the System for 10 AUs on 77 Sequences from the SAL Data Set for the LPQ-TOP in terms of frame-based AU detection accuracy: classification rate (CR), recall (RC), precision (PR), F1-measure (F1) and 2AFC. θ is the optimal window length for the LPQ-TOP descriptor, N the number of videos that contain an AU.

AU	θ	N	CR	RC	PR	F1	2AFC
1	15	8	90.4	40.88	48.79	44.48	0.78
2	19	10	88.6	46.87	58.76	52.15	0.93
6	11	28	91.2	97.7	93.2	95.4	0.85
7	15	7	55.9	34.7	71.8	46.79	0.78
10	15	13	91.7	98.5	59.8	74.42	0.86
12	19	35	94.00	94.00	100.00	96.90	1.00
23	11	6	52.79	68.00	42.40	52.23	0.60
25	5	33	82.40	82.40	100.00	90.35	1.00
26	15	18	75.80	82.52	86.81	84.60	0.64
45	15	17	60.30	45.12	71.51	55.31	0.68
AVG	-	-	73.31	69.06	73.31	69.26	0.81

due to the fact that non-frontal textures are significantly different than frontal ones. Training a pose-independent texture-based method requires a larger amount of training data (with examples of all AUs for all poses), which is challenging due to the lack of adequate databases and the cost of FACS annotation.

However, we still wanted to measure the performance of our method under these conditions. We followed the instructions of the FERA challenge [45], where the system was trained on the 87 training sequences provided, and tested on the 71 test sequences. The results of this experiment are shown in Table 4, where the baseline performance of the challenge is also shown. In this case, our LPQ-TOP-based system produces an increase of 4.9% respect to the baseline system.

As can be seen from Table 4, even for a database like GEMEP-FERA, where the alignment of the images is very challenging, the use of dynamic features is still beneficial. In fact, the relative performance increase when using dynamic features compared to when using static features is very similar. In particular, when using LPQ-TOP features instead of LPQ features the average performance, measured in 2AFC score, increases by 8% for the MMI database and an 8.5% for the GEMEP-FERA dataset. Another interesting aspect is that the average optimal window length is reduced (from an average optimal window length of 19 frames for MMI, to 14 frames for SAL, and 8 frames for GEMEP). This may be due to a greater challenge posed by alignment. The more images to be aligned, the more impact registration errors will have, as increasingly more noisy pattern variations will be included in learning.

5.2 Frame-based AU temporal segment detection

The performance of the AU temporal segment detection is evaluated on the MMI and the SAL datasets. Analogous to previous related studies, only sequences that have the target AUs activated are considered for testing.

For the case of the MMI database, the existence of annotations in terms of temporal segments allows per-

TABLE 4
2AFC of frame-based AU activation detection on the GEMEP-FERA test dataset. θ is the optimal window length for LPQ-TOP.

AU	baseline	LPQ	θ	LPQ-TOP
1	0.790	0.671	3	**0.846**
2	**0.767**	0.664	3	0.749
4	0.526	0.587	7	**0.639**
6	0.657	**0.673**	11	0.658
7	0.556	0.612	3	**0.629**
10	**0.597**	0.524	7	0.567
12	0.724	0.785	11	**0.827**
15	**0.563**	0.543	7	0.541
17	0.646	0.678	7	**0.713**
18	0.610	0.714	15	**0.715**
25	0.593	**0.607**	11	0.502
26	0.500	0.611	11	**0.709**
AVG	0.628	0.623		**0.677**

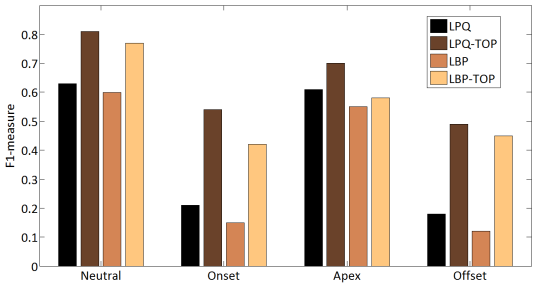

Fig. 12. Comparison of the F1 measures of temporal segment detection attained by using LBP-TOP and LPQ-TOP from the MMI database, measured per frame.

formance measures with respect to the above-described 4-class problem. Fig. 12 shows the average F1-measure attained when using LPQ-TOP, LBP-TOP and their static variants. On average, using LPQ-TOP attains higher accuracy than using LBP-TOP. Also the use of dynamic appearance features boosts the accuracy for temporal segment detection, especially for the onset and offset phases. This is unsurprising, as the appearance of an AU during the onset and associated offset frames is very similar. Thus, it will be very difficult for a classifier to distinguish them relying solely on static appearance.

TABLE 5
Percentages of early/on time/late detection per transition, tested on the MMI dataset

	Early	On time	Late
Neutral → Onset	22.65	47.64	29.71
Onset → Apex	19.63	23.92	56.45
Apex → Offset	20.37	53.23	26.41
Offset → Neutral	15.50	30.50	54.01

For the MMI database, we also examined the error in the estimated duration of the temporal segments and of the total AU episode. This results are shown in Fig. 11. For most AUs, the average error per temporal segment is less than 8 frames, and the prediction of the duration of the offset temporal segment usually has the largest error. Onset and apex error for AU45 and AU46R are missing due to their brevity (e.g., the average apex duration of

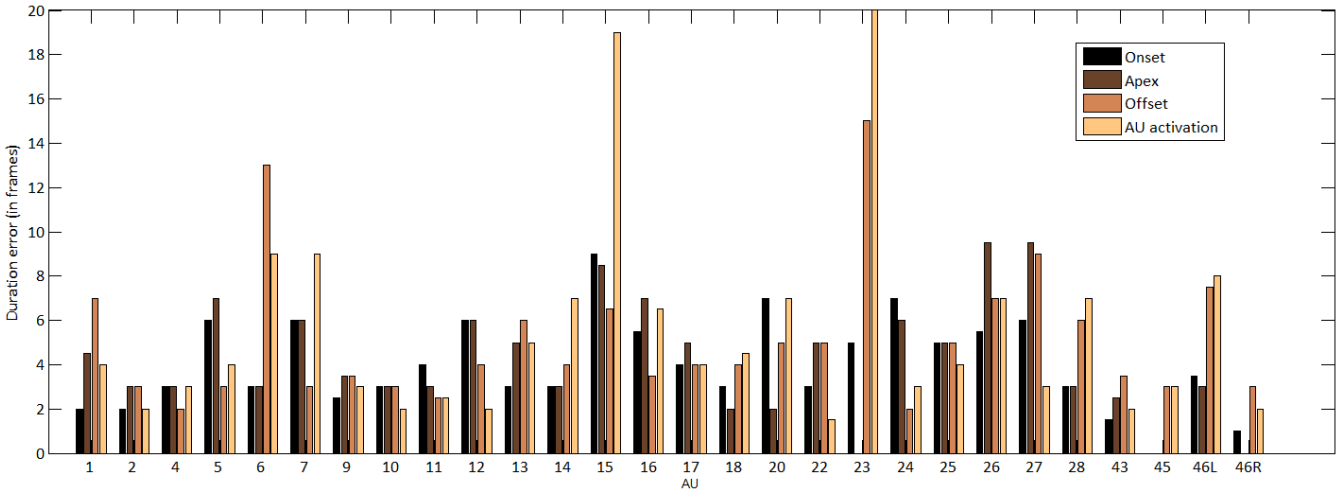

Fig. 11. Temporal segment (onset, apex, offset) duration error and the entire facial action duration error. Results are average per AU, and measured in frames and tested on the MMI dataset

AU45 is 1.2 frames). It is important to note that the error of the total AU activation duration is far less than the sum of the temporal-segment-duration errors. To wit, if the apex segment has been predicted to last too long, the offset phase will start late and will result in an error in the offset phase duration too, thus the error is effectively double counted.

TABLE 6
Confusion Matrix (percentages) for AU12 (Row represents true labels and column represents predictions)

	Neutral	Onset	Apex	Offset
Neutral	96.00	1.80	2.10	1.70
Onset	14.20	51.80	34.00	0.00
Apex	4.20	0.00	95.10	0.80
Offset	28.20	0.00	17.10	54.70

Table 5 presents the proportion of early, timely, and late detections for all correctly detected transitions. As we can see, there is a larger portion of predictions being late for the onset→apex and offset→neutral transitions. The early detection for the onset→apex and offset→neutral transitions can be illustrated on the example of AU12 (as shown in Table 6). We can see that both neutral and apex phases have very low confusion with any other class. Effectively, almost all errors are due to confusions between onset and apex, onset and neutral, offset and neutral, and offset and apex. This is logical as most AUs start and end in a very subtle manner, visible to the human eye but not sufficiently pronounced to be detected by an automatic method. Also note that there is no confusion between onset and offset either way.

For the SAL dataset, as described in Sec. 4.1, only sequences of 4 out of 10 subjects were annotated on a frame-by-frame basis in terms of AU temporal segments. The remaining sequences were annotated in a frame-by-frame manner only in terms of AU activation. Following the approach in [22], we used the former set for training and the latter for testing. As there is no temporal segment annotation in the test data, the prediction is converted into AU activation to compute the results.

TABLE 7
F1-measure classification accuracy of Hybrid approach for distinguishing the four temporal segments from the MMI database using LPQ-TOP. θ is the optimal window lengths for the LPQ-TOP descriptor. $F1_{act}$ is the F1-measure after converting into AU activation.

AU	θ	Neutral	Onset	Apex	Offset	$F1_{act}$
1	7	78.29	63.26	76.87	61.05	85.15
2	15	87.36	63.59	71.81	62.12	79.32
4	15	67.36	63.59	71.81	62.12	61.61
5	15	75.05	50.15	62.61	36.12	41.59
6	11	82.56	49.44	70.61	22.22	58.89
7	19	58.63	25.64	52.04	24.59	45.13
9	3	74.86	70.96	78.31	37.74	80.70
10	7	84.07	61.90	82.97	58.54	86.46
11	7	94.86	61.78	89.82	61.36	89.61
12	15	93.21	66.02	88.95	69.61	86.55
13	11	90.38	67.15	90.33	47.73	93.04
14	7	72.80	62.94	76.64	34.78	89.14
15	3	71.39	21.85	66.60	39.42	52.19
16	11	71.99	47.22	70.74	42.05	71.06
17	3	72.81	51.85	68.73	44.53	74.42
18	7	91.21	50.00	76.40	41.54	58.26
20	15	67.92	43.24	64.09	41.54	44.63
22	7	75.90	62.80	65.45	44.56	89.65
23	3	54.10	26.56	54.49	3.57	43.08
24	7	79.60	53.57	83.98	47.62	53.42
25	7	91.15	58.88	78.01	53.33	82.47
26	3	66.06	38.62	63.53	41.13	71.04
27	3	96.96	72.07	35.06	64.71	63.02
28	7	90.79	68.68	84.58	68.69	83.56
43	11	89.07	49.60	63.73	55.32	63.48
45	3	96.96	72.07	35.06	64.71	63.02
46L	3	95.44	34.81	55.94	66.06	50.73
46R	3	95.64	58.82	30.49	62.65	69.46
AVG	-	80.77	53.82	69.87	48.79	69.85

The results for the SAL dataset are given in Table 8. The obtained classification rate is 80.3% and an average F1-score is 79.3%. The poor performance (recognition rate) is reported for AU7, AU23 and AU45. The duration of AU45 is very short (average apex duration of 1.2 frames). The confusion between AU7 and AU45

TABLE 8
Results for Testing the LPQ-TOP-based method for 10 AUs on 77 Sequences from the SAL Dataset in terms of frame-by-frame AU activation detection. θ is the optimal window lengths for the LPQ-TOP descriptor.

AU	θ	NT	CR	RC	PR	F1	2AFC
1	15	8	90.85	61.90	64.36	63.10	0.71
2	15	10	95.71	86.89	58.56	69.97	0.87
6	11	28	98.50	98.93	99.55	99.24	0.57
7	5	7	42.33	41.62	100.00	58.77	0.75
10	5	13	98.86	100.00	98.86	99.43	0.60
12	5	35	97.18	97.18	100.00	98.57	1.00
23	11	6	58.29	64.85	68.16	66.46	0.60
25	5	33	83.54	83.54	100.00	91.03	1.00
26	7	18	81.61	84.26	96.30	89.88	0.65
45	7	17	55.75	43.61	80.23	56.51	0.63
AVG	-	-	80.26	76.28	86.60	79.30	0.74

is another reason for the poor performance. For AU23, in spontaneous expressions, the appearance changes are even more subtle than those shown in posed expressions.

5.3 Cross-database evaluation

In order to test the ability to generalise to novel conditions, a cross-database test was performed. Since there are no similar databases with available annotations in terms of temporal segments, we restricted the evaluation to the problem of AU activation detection. For posed facial expressions, the system is trained on the 264 sequences from the MMI and tested on the 55 sequences from the CK databases that met the restrictions in length imposed by the feature descriptor window length. For the cross-database experiment on spontaneous expressions, the system is trained on the 35 sequences from the SAL dataset which are fully annotated, and tested on the SEMAINE database. As only the sparse annotation of 10 sequences is available in the SEMAINE database, our system is evaluated only on the annotated frames.

Average results are shown in Table 9. The tests were run on those AUs available in both datasets using the optimal window size obtained from the trained database. From Table 9, we can see that the average result is, as expected, lower than the results for training and testing on the same dataset. The difference between the CK and MMI databases is partially explained by differences in annotation styles. For SEMAINE and SAL, the coding system is consistent, but aspects of the databases such as resolution, lighting conditions, codec artefacts, and camera positions differ. Note that some of these results are obtained with very little data.

TABLE 9
Cross-Database testing (Average 2AFC score over 15 AUs using LPQ-TOP)

TRAIN, TEST	CR	RC	PR	F1	2AFC
MMI, CK	80.56	50.30	51.94	51.11	0.80
MMI, MMI	94.70	63.80	77.80	68.40	0.87
SAL, SEMAINE	70.10	60.10	68.45	64.00	0.73
SAL, SAL	73.31	69.06	73.31	69.26	0.81

5.4 Comparison to Earlier Work

We compared our method to earlier works that reported results on the MMI, UNBC-McMaster pain, SAL and GEMEP-FERA datasets. Note that using the same database for testing does not necessarily mean that the evaluated methodologies were trained and tested in exactly the same experimental setup, e.g. on the same number of videos or using the same parameter optimisation strategy.

Table 10 gives an overview of the existing systems that report their performance in terms of frame-by-frame AU detection and temporal segment detection on the MMI database. In order to effectively compare the performance regarding AU temporal segment detection, we reported two quantitive measures. $F1_{bin}$ is the F1-measure after binarising the temporal segment results to effectively obtain AU activation coding, and $F1_{mean}$ is the average of the per-class F1-measures with respect to the temporal segments (see Table 7). For AU detection, we achieve an average F1-measure of 68.4% which clearly outperforms all the other systems.

For temporal segment detection, we compare with the FFD-based method [22], the work using geometric features [48] and the work of [36]. The results show that our system based on LPQ-TOP features outperforms [48] and [22]. Note that the classifiers and the training procedure used for AU detection here are exactly the same as those used in [48], and for the temporal analysis they are the same as in [22]. In this way the machine learning algorithm used in our system is not accountable for the superior performance, and thus it is possible to determine the relative merit of the LPQ-TOP dynamic appearance descriptor. By contrast, [36] attained 65% using LBP but with a more sophisticated temporal model Lap-CORF and the methodology has been applied on the upper-face AUs only. In principle, it is expected a higher score can be reached by using a combination of LPQ-TOP and Lap-CORF.

The results on the UNBC-McMaster pain database have been compared with those in [25]. The detailed results (2AFC) have been shown in Table 2. Again we have shown that the use of dynamic appearance features outperforms those of geometry-based and static appearance-based features for most AUs. Note that [19] has tested on the same database by using appearance features and regression. However, the performance of the algorithm is measure using mean square error and correlation coefficient, which cannot be directly compared to the 2AFC measure more commonly used. For the SAL dataset, only one other work reported results on it. The authors in [22] reported an average F1-measure of 75.52% for temporal segment detection (binarised results). Referring to Table 8, we can see that our system produces superior results of 79.30%. For comparisons with the existing systems on the GEMEP-FERA dataset please see [45]. Our system achieves an average 2AFC score of 0.677 which ranks third in the challenge despite

TABLE 10

Comparison of AU activation detection and temporal segment detection methods on the MMI Database. Nr is the number of sequences used, CR_{act}, $F1_{act}$ and $2AFC_{act}$ is the performance for frame-by-frame AU detection, $F1_{bin}$ is the F1 measure after binarising the temporal segments results, $F1_{mean}$ is the mean of F1 measure for difference segments.

System	feature type	classification method	AUs	Nr	CR_{act}	$F1_{act}$	$2AFC_{act}$	$F1_{bin}$	$F1_{mean}$
This work	Appearance	GentleBoost& MM	25	264	0.947	0.684	0.867	0.700	0.630
Valstar & Pantic 2012 [48]	Geometric	GentleSVM & MM	22	244	0.953	0.533	-	-	0.615
Rudovic et al. 2012 [36]	Appearance	Lap-KCORF	9	264	-	-	-	-	0.650
Jiang et al. 2011 [18]	Appearance	SVM	9	442	0.890	0.663	-	-	-
Koelstra et al. 2010 [22]	Appearance	GentleBoost MM	27	264	-	-	-	0.651	-
Valstar &Pantic 2007 [47]	Geometric	AdaBoost+SVM	23	196	-	0.660	-	-	-
Pantic & Patras 2005 [33]	Geometric	Rule-based	27	299	0.936	-	-	-	-

our method not being designed to deal with non-frontal poses.

6 Conclusions and Future Work

We have presented a novel approach to explicit analysis of the temporal dynamics of facial actions using the dynamic appearance descriptor LPQ-TOP. Extensive experimentation has shown that this dynamic appearance descriptor is highly suited for the problem of AU temporal segment detection, outperforming all previous works that reported on this task. The proposed methodology has also been shown to attain superior AU activation detection.

Given the descriptive power of LPQ-TOP in terms of capturing the dynamics of facial actions, it seems natural to extend our work to recognise more facial dynamic characteristics such as the intensity of AUs and their frequency of occurrence. In the light of our performance on the GEMEP-FERA dataset, it has become clear that head-pose invariant face registration is one obstacle that needs to be resolved in order for our approach to work on arbitrary data. Given the complementary nature of dynamic appearance and geometric features, fusing the two would also be a natural extension of this work.

Acknowledgments

This work has been funded in part by the European Community's 7th Framework Programme [FP7/20072013] under the grant agreement no 231287 (SSPNet). The work of Michel Valstar is supported by Horizon Digital Economy Research, RCUK grant EP/G065802/1. The work of Brais Martinez is also funded in part by the EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching. 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, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481, 2004.

[2] Z. Ambadar, J. Schooler, and J. F. Cohn. Deciphering the enigmatic face: The importance of facial dynamics to interpreting subtle facial expressions. *Psychological Science*, 2005.

[3] B. S. Arman Savran and M. T. Bilge. Regression-based intensity estimation of facial action units. *Image Vison Computing*, 2011. In press.

[4] 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.

[5] M. S. Bartlett, J. Larsen, J. C. Hager, and P. Ekman. Classifying facial actions. In *Advances in Neural Information Processing Systems*, pages 823–829, 1996.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[7] J. F. Cohn and P. Ekman. *Measuring facial actions*. Oxford University Press, 2005.

[8] 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, 2004.

[9] J. Delannoy and J. McDonald. Rankboost with L1 regularization for facial expression recognition and intensity estimation. In *Proc. IEEE Int. Conf. Computer Vision*, pages 1018–1025, 2009.

[10] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. Heylen. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In L. Devillers, J.-C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner, editors, *LREC Workshop on Corpora for Research on Emotion and Affect*, pages 1–4. ELRA, 2008.

[11] P. Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000:205–221, 2003.

[12] P. Ekman, W. Friesen, and J. C. Hager. Facial action coding system. In *A Human Face*. Salt Lake City, UT, 2002.

[13] P. Ekman and E. L. Ronsenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System*. Oxford University Press, 2005.

[14] D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

[15] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *Int'l Journal of Synthetic Emotion*, 1(1):68–99, 2010.

[16] H. Gunes and M. Piccardi. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(1):64–84, 2009.

[17] M. Hoque, L.-P. Morency, and R. W. Picard. Are you friendly or just polite? - analysis of smiles in spontaneous face-to-face interactions. In *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part I*, pages 135–144, 2011.

[18] B. Jiang, M. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, 2011.

[19] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *Proceedings of the International Symposium on Visual Computing*, Crete, Greece, July 2012.

[20] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[21] M. Kim and V. Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. In *European Conference on Computer Vision*, pages 649–662, 2010.

[22] 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.
[23] J. J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li. Automated facial expression recognition based on facs action units. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 390–395, 1998.
[24] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 41(3):664–674, 2010.
[25] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. W. Chew, and I. Matthews. Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database. 30(3):197–205, 2012.
[26] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 57–64, 2011.
[27] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *IEEE Workshop on CVPR for Human communicative Behavior analysis (CVPR4HB)*, 2009.
[28] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
[29] B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic. Local evidence aggregation for regression based facial point detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2012. (in press).
[30] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3:5–17, 2012.
[31] 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.
[32] M. Pantic. Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. *Philosophical Transactions of The Royal Society B: Biological sciences*, 365(1535):3505–3513, 2009.
[33] M. Pantic and I. Patras. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In *Proc. IEEE Int'l Conf. on Systems, Man and Cybernetics*, pages 3358–3363, 2005.
[34] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 36:433–449, 2006.
[35] M. Pantic, L. Rothkrantz, and H. Koppelaar. Automation of non-verbal communication of facial expressions. In *Proceedings of European Conf. Multimedia (EUROMEDIA '98)*, pages 86–93, 1998.
[36] O. Rudovic, V. Pavlovic, and M. Pantic. Kernel conditional ordinal random fields for temporal segmentation of facial action units. In *Proceedings of the 12th European Conference on Computer Vision (ECCV-W'12).*, 2012.
[37] O. Rudovic, V. Pavlovic, and M. Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
[38] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. Recognition of 3d facial expression dynamics. *Image and Vision Computing*, 2012. (in press).
[39] 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.
[40] C. Shan, S. Gong, and P. W. Mcowan. Dynamic facial expression recognition using a bayesian temporal manifold model. In *In Proc. British Machine Vision Conference*, pages 297–306, 2006.
[41] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.
[42] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.
[43] 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.
[44] M. Valstar, H. Gunes, and M. Pantic. How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of ACM Int'l Conf. Multimodal Interfaces*, pages 38–45, 2007.
[45] M. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *Systems, Man, and Cybernetics, part B*, 2012. in print.
[46] 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.
[47] 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.
[48] M. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 42(1):28–43, 2012.
[49] M. Valstar, M. Pantic, Z. Ambadar, and J. Cohn. Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In *Proceedings of ACM Int'l Conf. Multimodal Interfaces*, pages 162–170, 2006.
[50] M. F. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *Proceedings of IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 2729–2736, 2010.
[51] A. C. Williams. Facial expression of pain: An evolutionary account. *Behavioral and Brain Sciences*, 25(4):439–488, 2002.
[52] 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.
[53] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005.
[54] 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.

Bihan Jiang received her MSc degree in Computer Science from Imperial College London in 2009. She is currently pursuing her PhD degree with the intelligent Behaviour Understanding Group (iBUG) under the supervision of Prof. Maja Pantic. Her areas of interest include applying machine learning and pattern recognition approaches to the study of human (social) behaviour and to build better human interfaces.

Michel F. Valstar (M'09) is a Lecturer in the Mixed Reality Lab at the University of Nottingham. He received his masters degree in Electrical Engineering at Delft University of Technology in 2005 and his PhD in computer science with the intelligent Behaviour Understanding Group (iBUG) at Imperial College London in 2008. Currently he is working in the fields of computer vision and pattern recognition, where his main interest is in automatic recognition of human behaviour. In 2011 he was the main organiser of the first facial expression recognition challenge, FERA 2011. In 2007 he won the BCS British Machine Intelligence Prize for part of his PhD work. He has published technical papers at authoritative conferences including CVPR, ICCV and SMC-B and his work has received popular press coverage in New Scientist and on BBC Radio. He is also a reviewer for many journals in the field, including Transactions on Pattern Analysis and Machine Intelligence, Transactions on Affective Computing, Systems, Man and Cybernetics-B and the Image and Vision Computing journal.

Brais Martinez (M'10) received his BC degree in Mathematics from the Universidad de Santiago de Compostela in 2003 and a MD and PhD in Computer Science from the Universitat Autonoma de Barcelona in 2006 and 2010 respectively, under the supervision of Xavier Binefa. He is currently affiliated as a research associate with the intelligent Behaviour Understanding group (iBUG) of Maja Pantic at Imperial College London.

Maja Pantic (M'98, SM'06, F'12) is Professor in Affective and Behavioural Computing at Imperial College London, Department of Computing, UK, and at the University of Twente, Department of Computer Science, the Netherlands. She received various awards for her work on automatic analysis of human behaviour including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She currently serves as the Editor in Chief of Image and Vision Computing Journal and as an Associate Editor for both the IEEE Transactions on Systems, Man, and Cybernetics Part B and the IEEE Transactions on Pattern Analysis and Machine Intelligence. She is an IEEE Fellow.