Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition

Timur R. Almaev
The University of Nottingham
psxta4@nottingham.ac.uk

Michel F. Valstar
The University of Nottingham
michel.valstar@nottingham.ac.uk

Abstract—**Facial actions cause local appearance changes over time, and thus dynamic texture descriptors should inherently be more suitable for facial action detection than their static variants. In this paper we propose the novel dynamic appearance descriptor Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP), combining the previous success of LGBP-based expression recognition with TOP extensions of other descriptors. LGBP-TOP combines spatial and dynamic texture analysis with Gabor filtering to achieve unprecedented levels of recognition accuracy in real-time. While TOP features risk being sensitive to misalignment of consecutive face images, a rigorous analysis of the descriptor shows the relative robustness of LGBP-TOP to face registration errors caused by errors in rotational alignment. Experiments on the MMI Facial Expression and Cohn-Kanade databases show that for the problem of FACS Action Unit detection, LGBP-TOP outperforms both its static variant LGBP and the related dynamic appearance descriptor LBP-TOP.**

I. Introduction

The face is the primary means to identify other members, to determine their personality, to interpret what has been said, and to understand someone's emotional state and intentions on the basis of the shown facial expression. Automatic face analysis in general and facial expression recognition in particular has therefore become a popular topic in recent years, with the aim of creating machines with interfaces that are better aligned to human communication. One of the biggest challenges in creating such interfaces lies in the fact that people communicate a large amount of information non-verbally through e.g. body gestures and facial expressions [1], which are difficult to measure automatically.

In early 1978 Paul Ekman and Wallace Friesen published their adopted version of Facial Action Coding System (FACS) [2], originally developed by the Swedish anatomist Carl Herman Hjortsjo in the late 1960s [3]. FACS is the best known and most commonly used system developed for human observers to objectively describe facial activities. The coding system defines 32 atomic facial muscle actions called Action Units (AUs). With FACS, every possible facial expression can be described as a combination of AUs. It enabled social scientists and psychologists to use mathematical and statistical theory to study facial expressions.

In addition, because AUs are independent of interpretation, they can be used as an objective, low-dimensional basis for assigning meaning to expressions, including the basic emotions and cognitive states like interest and puzzlement, social behaviours like agreement and disagreement, social signals like status, trustworthiness and so on. For instance, an expression typically associated with happiness contains AU6 and AU12 and sadness contains AU1, AU4 and AU15. Researchers have employed FACS to study everything from deception detection [4], [5], [6] to data-driven avatars [7].

There are a number of existing automatic facial expression recognition techniques available today, but there remain a number of outstanding issues in this field. These include the ability to deal with non-frontal head poses, accurate and explicit temporal analysis of facial actions, subject independent facial feature tracking, true intensity estimation, and dealing with subtle, natural expressions. In this paper we will propose a novel appearance descriptor designed to address the issues of temporal analysis and intensity estimation.

While the recently held first Facial Expression Recognition Challenge (FERA) indicated that the recognition of a small set of discrete expressions, such as the six basic emotions [8], is basically a solved case if the user is known, the same if far from true for AU detection. In particular, the winners of that challenge attained an average 2AFC score of only 75.2% [9]. There is thus still a large gap between accurate discrete expression recognition and AU detection, even on data that is recorded under fairly well-controlled conditions.

Almost every existing appearance-based approach to AU detection uses static descriptors, i.e. every data point/feature is defined based on a single moment in time. However, in essence AU detection is *action* detection, and rather than looking only at the current appearance, one should look at the *changes* in appearance *over time*. It is in this context that we present our novel approach to AU detection, aiming to improve accuracy while remaining subject independent and maintaining soft-real time computational performance. We propose to extend the previously successful LGBP method [9] to the temporal domain, resulting in Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP). We experimentally show that the proposed descriptor significantly outperforms its static variant LGBP and its close spatio-temporal relative LBP-TOP.

LGBP-TOP applies LGBP independently on three orthogonal planes: the spatial x-y plane, and the temporal x-t and y-t planes. In the temporal planes, it essentially encodes facial movements as up/down or left/right movements of edges. Of course, for facial expression analysis, it is essential that the motion of these edges is caused by facial muscle activations, rather than by rigid head movements, camera motion, or errors in the detection and registration of the face.

As face registration aims to remove any motion caused by head and/or camera motion, it is the errors in face registration that are important to take into account when evaluating a dynamic appearance descriptor. In this paper, we go beyond simply extending yet another appearance descriptor to the temporal domain. We additionally hypothesise that LGBP is more robust to face registration errors than simple LBP is, due to the smoothing effect of the Gabor filters. We experimentally validate this hypothesis, comparing LGBP-TOP with LBP-TOP on data where noise is added in the form of random rotation registration errors.

Our contribution is thus threefold:

- We propose the dynamic appearance descriptor Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP)
- We apply LGBP-TOP for the first time to AU detection and show a significantly improved accuracy over all related approaches (i.e. LBP, LBP-TOP and LGBP)
- We analyse the robustness of TOP methods to variations in rotation errors during face alignment and show that LGBP-TOP is least sensitive to this

The remainder of this paper is organised as follows: section II provides an overview of the related work; section III gives a detailed theoretical description of the proposed approach; section IV provides all the methodological details about the system implemented and data used to evaluate the proposed approach; and section V presents the evaluation results. Finally, our concluding remarks are provided in section VI.

II. Related Work

Facial expression recognition in general and Action Unit Detection in particular has been studied extensively in the past decade. As a result, it is impossible to provide a comprehensive review of the field here. Instead we provide an overview of the relevant works only, focussing on related static and dynamic appearance-based methods. For a general overview of the field of expression recognition we refer the reader to two excellent recent surveys [10], [11].

Gabor wavelet filtering is a successful filter bank approach that is sensitive to fine wave-like image structures such as those corresponding to wrinkles and bulges. In order to capture these structures, it is important to use banks of filters with the right frequencies given a certain face resolution so that the wave frequency corresponds to the expected wrinkle characteristics. In addition, the right filter orientations have to be included to capture relevant directions of edges, and one or more spatial scales have to be selected to determine their spatial extent. Gabor magnitudes are commonly adopted as features as they are robust to misalignment (e.g. [12], [13], [14]). Computing Gabor filters has however a high computational cost, and the dimensionality of the output can be very large, especially if they are applied holistically with a wide range of frequencies, scales, and orientations.

The LBP features were originally proposed for texture analysis, and recently have become very popular for face analysis. The local binary pattern of a pixel is defined as a 8-dimensional binary vector that results from comparing its intensity against the intensity of each of the neighbouring pixels (see Fig. 1). The LBP descriptor is a histogram where each bin corresponds to one of the different possible binary patterns, resulting in a 256-dimensional descriptor of the basic LBP. However, it has been shown that some of the patterns are more prone to encode noise or spurious structures. In practise, the most popular version of LBPs is the so-called uniform pattern LBP [15].

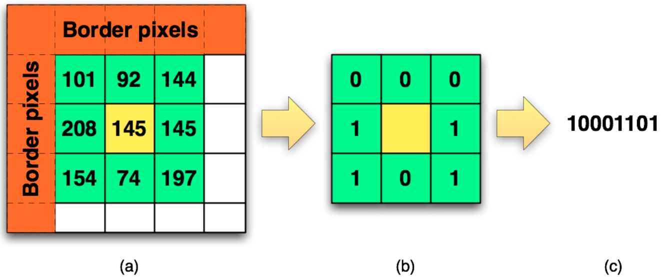

Fig. 1. Computing LBP response from a pixel's local neighbourhood.

The main advantages of LBP features are their tolerance to illumination changes, their computational simplicity, and their sensitivity to local structures while remaining robust to variations in face alignment [16]. They are however not inherently robust to rotations, though some effort has been made to modify the descriptor to include this property [15], but in practice this extension was less accurate than the original. A review of other LBP-based descriptors can be found in [17].

Recently, Senechal et al. [9] and Wu et al. [18] proposed to use appearance features extracted in a multi-layer architecture, of which Local Gabor Binary Patterns (*LGBP*) is a prime example. LGBPs are extracted by first creating a set of Gabor magnitude response images (one for every filter in a filter bank) and then applying an LBP operator to each of them. This has been shown to be very robust to illumination changes and misalignment [19]. The winner of the FERA 2011 AU detection sub-challenge adopted this architecture. Wu et al. [18] employed two layers of Gabor features, called G^2. The idea was that such representation could encode image textures that go beyond edges and bars. They further compared single layer (LBP, Gabor) and dual layer (G^2, LGBP) architectures for texture-based AU detection, and concluded that dual layer architectures provide small but consistent improvements.

Zhao and Pietikainen [20] proposed an extension of LBPs to spatio-temporal volumes. In order to make the approach computationally simple, the proposed extension computes LBP features only on Three Orthogonal Planes (TOP) within a fixed temporal window: XY, XT, and YT, resulting in the *LBP-TOP* descriptor (as shown in Fig.3). Similarly, LPQ is extended to *LPQ-TOP* [21]. In [21], the performance of both LBP-TOP and LPQ-TOP for automatic AU detection was evaluated and compared to that of their static counterparts. The authors apply the descriptors in a block-based holistic manner, and conclude that dynamic appearance descriptors offer a significant performance improvement.

In [22] the authors proposed dynamic features based on Haar-like features. During a training phase, the distribution of values of each Haar-like feature under an AU is modelled through a Normal distribution. To build the dynamic feature, the full set of Haar-like features for every frame within a

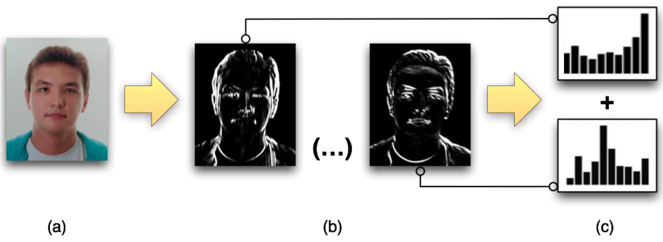

Fig. 2. Creating LGBP features: a) original image, b) Gabor pictures, and c) concatenation of resulting histograms after applying LBP.

temporal window is computed. Then, a binary pattern of the length of the temporal window is obtained for each feature by thresholding over the Mahalanobis distance respect to the corresponding Normal distribution. This has been extended in [23], although it was only applied to the recognition of prototypical facial expressions.

III. LGBP-TOP

Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) splits face video into space-time video-volumes, convolves all x-y, x-t, and y-t intersections of each video with a bank of Gabor filters, and finally applies LBP to all of the filtered intersections. Below we describe this process in detail.

A set of Gabor filters results in a number of filtered copies of the original image, commonly called Gabor pictures (GP). Each GP is the result of the original image convolved with the 2D complex Gabor function. A 2D complex Gabor function in turn is the convolution of a 2D sinusoidal carrier with spatial frequencies u_0 and v_0 and a 2D Gaussian with amplitude K, orientation θ and spatial scales a and b. Here, for simplicity we employ the same spatial scales and frequencies in x and y direction, and unit amplitude, i.e. $a = b = \sigma$, $u_0 = v_0 = \phi$, $K = 1$. Our 2D complex Gabor function then becomes:

$$\begin{aligned} G(x, y) = &\exp(-\pi\sigma^2((x - x_0)_r^2 + (y - y_0)_r^2) \\ &\exp(j(2\pi\phi(x + y) + P))) \end{aligned} \tag{1}$$

$$\begin{aligned} (x - x_0)_r &= (x - x_0)\cos\theta + (y - y_0)\sin\theta \\ (y - y_0)_r &= -(x - x_0)\sin\theta + (y - y_0)\cos\theta \end{aligned}$$

We take the magnitude response of this function, which cancels out the effect of the phase P. Figure 2 provides an example of how an original image results in a set of Gabor Pictures after being convolved with a set of Gabor filters.

In basic LBP, for every pixel of the image, its grayscale value is compared with those of the eight surrounding pixels (see Fig. 1). The value of each neighbour is set to 0 if its grayscale value is smaller than the central pixel's value and to 1 otherwise:

$$LBP(p_c) = \sum_{k=0}^{7} \delta(f_k - f_p)2^k, \tag{2}$$

$$\delta(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$$

Thus an eight digit binary number called the local binary pattern is composed, which allows 256 possible values.

It has been shown that patterns which contain at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is traversed circularly account for about 90% of the LBP responses, with only 59 patterns representing only strong edges and corners. This fact was used to create a modification of the original LBP operator called Uniform LBP, which allows to significantly reduce the amount of bins of the feature histogram by omitting all non-uniform patterns.

To lower the dimensionality of the problem further and attain a degree of shift-robustness, images are usually divided by a regular grid of $m \times n$ local regions from which LBP histograms can be extracted, which can then be concatenated into a histogram. This block-based representation of local texture descriptors was first proposed by Ahonen et al. [24].

In LGBP, the Gabor and LBP filtering follow one after the other. But whereas in normal LBP the LBP filter operates only on the original image, in LGBP it operates on a number of filtered images, because a bank of different Gabor filters is used to generate response images. The final LGBP feature histogram of the image is then composed by a simple concatenation of the histograms composed for each GP, with histogram blocks in the same manner as for LBP. For example, if we have 18 Gabor filters and apply Uniform LBP to an image split into 16 (4×4 grid) local regions we would obtain a feature histogram descriptor of dimensionality $18 * 16 * 59 = 16\,992$.

Volume Local Binary Patterns (VLBP) or 3D Local Binary Patterns (3D LBP) is an extension of the original LBP operator to the spatio-temporal domain[1]. VLBP considers a block of video frames as a single 3 dimensional array of grayscale values. Following the rules of the basic LBP for each pixel it calculates a binary number based on the intensity comparison between the centre pixel and all its neighbours, including the 9 neighbour pixels in the previous and next frame, which leads to $3 * 8 + 2 = 26$ long local binary patterns. This results in a VLBP dimensionality of 2^{26}, which makes VLBP calculation computationally hard and risks succumbing to the curse of dimensionality.

A simplified, more practical version of the approach was proposed by its creators to make it more attractive for further usage called Local Binary Patterns from Three Orthogonal Planes (LBP-TOP). LBP-TOP applies LBP on every xy, xt, and yt slice separately, averages the histograms over all slices in a single plane orientation, and concatenates the resulting histograms of the three dimensions. This is shown in Fig. 3. With LBP-TOP it is possible to combine motion and appearance analyse in one operator with the feature histogram length $3 * 2^8$ ($3 * 59$ if Uniform LBP is used).

In LGBP-TOP a set of GPs of each block of frames within a video sequence is created. Each GP-video-volume is created by filtering every frame within the block with a specific Gabor filter. These GP-video-volumes are consequentially processed by the LBP-TOP operator described above. The resulting feature histogram of the block is then composed by concatenation of histograms built for each of them.

[1] Note that despite its name, V-LGBP [25] does not encode temporal dynamics. Instead it operates on static images, treating the set of Gabor filter responses as the third dimension.

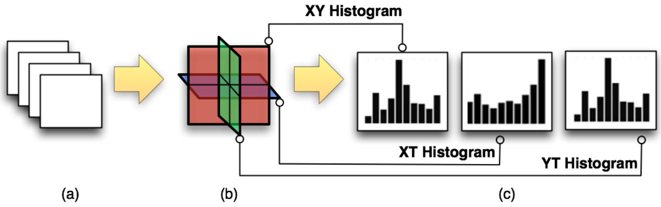

Fig. 3. Three planes in spatio-temporal domain to extract TOP features, and the histogram concatenated from three planes. a) original image, b) the x-y, y-t, and x-t planes, and c) concatenation of resulting histograms into a single feature set.

IV. Evaluation Methodology

Our aim is to ascertain that the spatio-temporal descriptor LGBP-TOP results in more accurate AU predictions than both its static counterpart LGBP and other dynamic appearance descriptors, in particular LBP-TOP. To do so, we used 314 videos of 17 different subjects from the MMI facial expression database [26] and from the original Cohn-Kanade database [27] 205 video sequences of 138 different subjects. While we acknowledge that these databases of mainly posed facial expressions do not necessarily allow generalisation to performance of a system in-the-wild, they are entirely suitable for the type of comparative studies we aim to do here. In this study, Support Vector Machines (SVM) have been used in a 10-fold subject independent cross-validation procedure.

Since MMI and Cohn-Kanade have different ways of organising data we applied different data preprocessing to each to normalise the data and prepare it for further processing with a variety of descriptors and SVMs. In MMI every expression is recorded in a single video file recorded at 25 frames per second and a spatial resolution of 720×576 pixels. In Cohn-Kanade all expressions are stored as sequences of images with a resolution of 640×490 pixels. Facial points were automatically detected in faces of both databases using the publicly available Local Evidence Aggregated Regression (LEAR) detector [28], which have been used in order to perform accurate face localisation.

Although the LEAR-detected points allow accurate face localisation, without adjustment the obtained facial image dimensions might vary (about 350×400 for Cohn-Kanade and about 200×300 for MMI). This variation is too small to seriously affect static descriptors, but might have very serious impact on dynamic approaches, where pixel to pixel precision is required within blocks of frames. Thus, for every block of frames composed, the following normalisation scheme was applied:

- Determine the smallest frame resolution within the block ($N \times M$);
- For every other frame only keep $N \times M$ pixels from its centre;

Note that we did not scale the image - as the temporal windows are relatively short, we assume here that variations in detected face dimensions are caused by errors in face localisation rather than real face motion.

Because of the large imbalance between positive and negative examples in AU detection, we employed the 2AFC classification performance measure to compare the various approaches. The 2AFC score is a good approximation of the area under the receiver operator characteristic curve (AUC). And while some previous works employ the F1 measure for the same reason, the 2AFC classification performance score considers both correctly identified positive as well as negative samples, whereas the F1 measure does not take True Negative predictions into account. In this particular study the 2AFC score has been calculated based on the SVM decision function output values as follows:

$$2AFC(\hat{Y}) = \sum_{i=0}^{n}\sum_{j=0}^{p}\sigma(P_j, N_i)\frac{1}{n \times p}, \tag{3}$$

$$\sigma(X,Y) = \begin{cases} 1, & \text{if } X > Y \\ 0.5, & \text{if } X == Y \\ 0, & \text{if } X < Y \end{cases}$$

where $\hat{Y}$ is a vector of decision function output values, n is the total number of true negative and p the total number of true positive instances in $\hat{Y}$, and P and N are subsets of $\hat{Y}$ corresponding to all positive and negative instances, respectively.

In order to focus on the performance of the features rather than the classifier, we used a linear kernel SVM. This requires only the slack variable C to be optimised. This value was determined by a separate automatic 3-fold cross-validation loop during the training phase of every fold, for every AU, from a range of 0.01 to 5000.

The actual Gabor filters used in the experimental system have been generated using the following set of values for the sinusoidal spatial frequency ϕ and Gaussian orientation θ:

$$\phi = \left(\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{8}\right) \tag{4}$$

$$\theta = \left(\frac{k\pi}{6}, k \in \{0...5\}\right) \tag{5}$$

This results in 18 different filters applied to all input frames. Frames were divided into 4×4 spatial blocks, and the temporal window length of the space-time video-volumes was fixed at 5 frames. This results in $18 \times 4 \times 4 \times 59 \times 3 = 50\,976$ features per set of frames. This is a very large number, and much larger than the number of training instances in our dataset. This risks a negative impact of the curse of dimensionality, and we thus employ feature selection to reduce the dimensionality of our classification problem. We used the feature selection technique proposed by Chen and Lin [29]. This involves calculating the F1-score for every feature, and then gradually adding 5% of the features with the highest score in a 5-fold subject independent cross-validation procedure, comparing at each stage the 2AFC score for the selected subset of features. The number of features increase until the 2AFC stops improving. The subset of features used to obtain the highest 2AFC is then used for the following experimental stages. The feature selection algorithm was applied to all descriptors being studied without any changes.

SVM training and testing is a very time consuming process, therefore we only tested AU detection performance on the following set: AU2, AU6, AU7, AU12, AU20, AU25, and AU45.

V. Experimental Results

This section provides an explicit explanation of the experiments performed on the LGBP-TOP approach being proposed as well as a number of similar techniques for the purposes of comparison. Two sets of experiments have been performed in order to check whether the hypotheses formulated in the introduction are correct or not.

Our first set of experiments is designed to determine whether LGBP-TOP provides better overall recognition performance than the approaches LBP, LGBP and LBP-TOP. The results are shown in Fig. 4 and Fig. 5, from which we can conclude the following:

- In general a combination of LBPs and a number of Gabor filters is more accurate than that of the original LBP
- Extending the appearance descriptors to the temporal domain using TOP improves accuracy
- LGBP-TOP attains the best results because of the combination of the above approaches

On the whole, LGBP-TOP has an improvement compared to LBP in terms of 2AFC score of 14% for the MMI data, and 18% for the Cohn-Kanade data, which is a considerable improvement. On the MMI data, AU2, AU6, and AU20 are detected with more than 90% 2AFC, and on Cohn-Kanade AU2, AU6, AU12, AU30, and AU25 are detected with over 93% 2AFC.

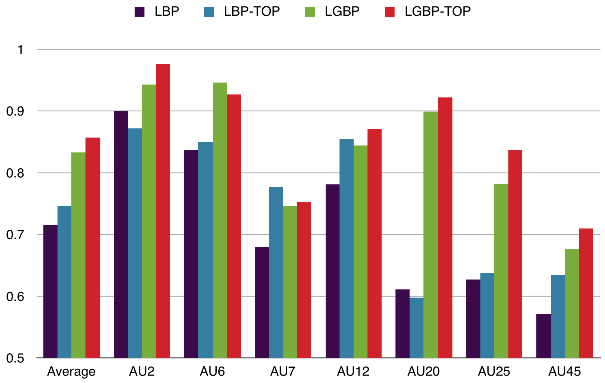

Fig. 4. AU detection results measured in 2AFC comparing LBP, LGBP, LBP-TOP and LGBP-TOP descriptors based on posed data taken from the MMI database

The second set of experiments is intended to show that Gabor filtering helps against artefacts caused by inaccurate face alignment, which is a very common situation in real-life settings. To verify this, for every space-time video-volume all but the first frame was perturbed by a random rotation around the centre of the image. We experimented with a random rotation drawn from a Normal distribution with mean 0 and a standard deviation of 3, 7 or 11 degrees. Results shown in Fig. 6 are the average 2AFC scores for AU2, AU6, and AU12.

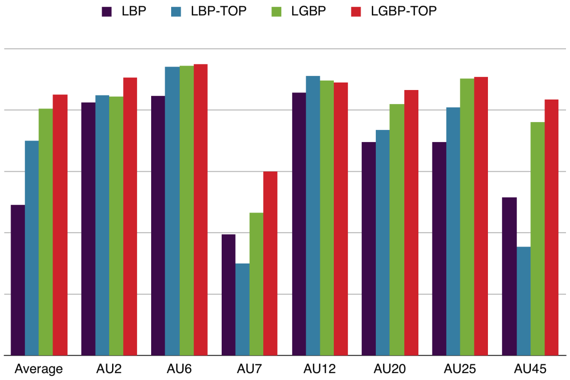

Fig. 5. AU detection results measured in 2AFC comparing LBP, LGBP, LBP-TOP and LGBP-TOP descriptors based on posed data taken from the Cohn-Kanade database

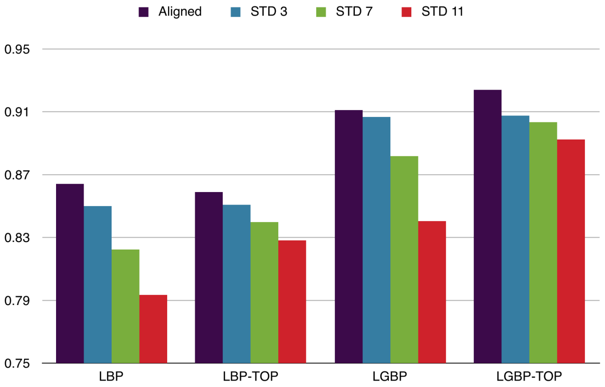

Fig. 6. Analysis of sensitivity to errors in alignment. Images are rotated randomly from a Normal distribution with std 3, 7 and 11 degrees. Accuracy measured in 2AFC.

As can be clearly seen from the obtained results, performance reduction for LBP-TOP is close to linear and therefore much more intense than that for LGBP-TOP, which shows a curve close to inverse logarithmic, confirming our hypothesis that the Gabor filtering reduces the sensitivity to registration errors. But perhaps the most unexpected result of this study is the observation that the static descriptors suffer far more from the registration errors than the TOP features. This is something that warrants further investigation.

Taking a closer look at the consistency of the selected features, we found that the number of temporal features selected was approximately twice as big as the number of static appearance features across all AUs ($60-65\%$ temporal). Considering however that TOP processing results in twice the number of temporal features in the original set, we can only conclude that both appearance and temporal features are selected with equal probability.

VI. Conclusion

Dynamic features are more powerful descriptors than their static counterparts, as they are usually a generalisation. Furthermore, they can truly encode an action, which has by definition a temporal component. However, the dimensionality of the resulting feature vector is very large, and using a fixed-length window is not a natural way of dealing with actions of varying speed. A novel dynamic appearance descriptor for automatic facial expression recognition called LGBP-TOP has been proposed, and shown to increase the overall level of recognition accuracy. The use of Gabor filters has also been shown to be moderately robust to face alignment errors in terms of random rotation errors of faces. A number of experiments show that LGBP-TOP is a very promising approach, and clearly shows that the right feature extraction is a very important aspect of facial expression recognition, and the possibilities of further improvement of the current technology should be carefully investigated. In particular we aim to investigate the performance on non-posed datasets for all AUs, and perform a sensitivity analysis to other alignment errors such as variations in shift and zoom levels. We also aim to address the issue of fixed-length temporal windows and variable frame rates.

Acknowledgements

The research was in part funded by the National Institute for Health MindTech Healthcare Technology Co-operative. The work of Michel Valstar is also supported by Horizon Digital Economy Research, RCUK grant EP/G065802/1.

References

[1] M. Pantic, R. Cowie, F. D'Errico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder, and A. Vinciarelli, "Social signal processing: The research agenda," in *Visual Analysis of Humans*. Springer Verlag, 2011, pp. 511–538.

[2] P. Ekman, W. V. Friesen, and J. C. Hager, *FACS Manual*. Network Information Research Corporation, May 2002.

[3] C. Hjortsjö, *Man's Face and Mimic Language*. Studentlitteratur, 1969.

[4] P. EKMAN, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003. [Online]. Available: http://dx.doi.org/10.1196/annals.1280.010

[5] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn, "Spontaneous vs. posed facial behavior: automatic analysis of brow actions," in *Proceedings of the 8th international conference on Multimodal interfaces*. ACM, 2006, pp. 162–170.

[6] M. F. Valstar, H. Gunes, and M. Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 38–45.

[7] E. B. Roesch, L. Tamarit, L. Reveret, D. Grandjean, D. Sander, and K. R. Scherer, "Facsgen: A tool to synthesize emotional facial expressions through systematic manipulation of facial action units," *Journal of Nonverbal Behavior*, vol. 35, no. 1, pp. 1–16, 2011.

[8] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: categories, origins, usage, and coding," *Semiotica*, vol. 1, pp. 49–98, 1969.

[9] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial Action Recognition Combining Heterogeneous Features via Multikernel Learning," *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 993–1005, 2012.

[10] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[11] Z. Zeng, M. Pantic, G. I. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[12] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, pp. 22–35, 2006.

[13] M. H. Mahoor, M. Zhou, K. L. Veon, M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," in *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, 2011, pp. 336–342.

[14] A. Savran, B. Sankura, and M. T. Bilge, "Regression-based intensity estimation of facial action units," *Image Vison Computing*, 2011, in press.

[15] 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*, vol. 24, no. 7, pp. 971–987, 2002.

[16] C. Shan, S. Gong, and P. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2008.

[17] D. Huang, C. Shan, and M. Ardabilian, "Local binary pattern and its application to facial image analysis: A survey," *IEEE Trans. Systems, Man and Cybernetics, Part C*, vol. 41, pp. 1–17, 2011.

[18] T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S.Bartlett, and J. R. Movellan, "Multi-layer architectures of facial action unit recognition," *IEEE Trans. Systems, Man and Cybernetics, Part B*, 2012, in print.

[19] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Combining aam coefficients with lgbp histograms in the multi-kernel svm framework to detect facial action units," in *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, 2011, pp. 860–865.

[20] G. Y. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary pattern with an application to facial expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 2, no. 6, pp. 915–928, 2007.

[21] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, 2011.

[22] P. Yang, Q. Liua, and D. Metaxas, "Boosting encoded dynamic features for facial expression recognition," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 132–139, 2009.

[23] ——, "Dynamic soft encoded patterns for facial event analysis," *Comp. Vision, and Image Understanding*, vol. 115, no. 3, pp. 456–465, 2011.

[24] T. Ahonen, A. Hadid, and M. Pietikainen, "Face recognition with local binary patterns," in *European Conference on Computer Vision*, 2004, pp. 469–481.

[25] S. Xie, S. Shan, X. Chen, and W. Gao, "V-LGBP: Volume based local Gabor binary patterns for face representation and recognition," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1–4.

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

[27] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 46–53.

[28] B. Martinez, M. Valstar, X. Binefa, and M. Pantic, "Local Evidence Aggregation for Regression Based Facial Point Detection," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1149–1163, 2013.

[29] Y.-W. Chen and C.-J. Lin, "Combining svms with various feature selection strategies," in *Feature Extraction*. Springer Berlin Heidelberg, 2006, pp. 315–324.