Improved Local Binary Pattern Based Action Unit Detection Using Morphological and Bilateral Filters

Anıl Yüce[1], Matteo Sorci[2] and Jean-Philippe Thiran[1]

[1]Signal Processing Laboratory (LTS5)
École Polytechnique Fédérale de Lausanne,
Switzerland
{anil.yuce;jean-philippe.thiran}@epfl.ch

[2]nViso SA
Lausanne, Switzerland
matteo.sorci@nviso.ch

Abstract— **Automatic facial action unit (AU) detection in videos is the key ingredient to all systems that utilize a subject face for either interaction or analysis purposes. With the ever growing range of possible applications, achieving a high accuracy in the simplest possible manner gains even more importance. In this paper, we present new features obtained by applying local binary patterns to images processed by morphological and bilateral filters. We use as features the variations of these patterns between the expressive and neutral faces, and show that we can gain a considerable amount of accuracy increase by simply applying these fundamental image processing tools and choosing the right way of representing the patterns. We also use these features in conjunction with additional features based on facial point geometrical relations between frames and achieve detection rates higher than methods previously proposed, using a small number of features and basic support vector machine classification.**

I. INTRODUCTION

From new generation game consoles to market research or software used for the treatment of psychopathologies, many applications and devices nowadays make use of facial analysis of users, consumers or patients. Automated facial action detection and classification therefore continues to be an important research issue in the computer vision area. The Facial Action Coding System (FACS) [1] is an objective and efficient way of describing any possible movement on the face using Action Units (AU), each of which define a certain movement of the facial muscles. The FACS has been extensively used by researchers in emotion as well as in human computer interaction systems. Accurate detection of the AUs in a simple and robust fashion is a very important step in facial analysis systems and in this paper we present a new method for extracting features from the facial texture that are able to efficiently describe facial actions.

Many automatic action unit systems have been proposed in the last decade using various methods for feature extraction and classification. One of the most well-known works is the one proposed by Bartlett et al. [2] where the authors use Gabor Features and Adaboost classifiers to detect the AU present in an image. The work by Valstar and Pantic [3] uses geometrical features instead and combines the approach with a GentleBoost feature selection algorithm and support vector machine classifiers. The fact that they use the transition of the facial points and their inter relations throughout a sequence allows them to perform detection of the temporal phases of AUs in addition to their presence. The more recently conducted Facial Expression Recognition Challenge [4], however, has shown that the best AU accuracy was achieved using geometric and texture features in combination[5]. In [5] the authors use Local Gabor Binary Pattern (LGBP) histograms and Active Appearance Model (AAM) features together in a multi-kernel SVM framework and achieve very high detection results.

Although the LBP and its many variants have been extensively investigated for AU and expression recognition purposes [6], too few of the works have gone further than extracting histograms on a fixed grid in 2D or 3D (the third dimension being time). In [7] the authors have successfully used the difference of the LGBP histograms between the neutral image and the peak expression. We adopt a similar approach, however we compute the LBP histograms obtained from overlapping windows and compute a single feature per window, which is the χ^2 distance between the histograms, resulting in a smaller number of features which search more extensively throughout the image. In addition, we apply three different filters (using morphology by reconstruction and bilateral filters) separately before applying the LBP transform on the image. This lets us obtain three different LBP transforms which define more clearly the edges than directly applying the LBP transform, and we show with experimental results that indeed the new features proposed achieve a better accuracy. We also show that combining these texture features with certain shape features we can achieve detection performances higher than other methods that have reported results on the same database that we use for our tests.

The rest of the paper is organized as follows: Section II describes the shape features, preprocessing methods and texture feature extraction procedure along with the feature selection and classification method that we use. In Section III we present the results obtained using texture features by themselves and in conjunction with shape features and compare these results to other methods. Finally, Section IV concludes the paper with a summary and outlook.

II. Proposed Method

In this section we explain in detail the method proposed for the AU detection system. Since our main contribution is in features extraction, the emphasis is also given to this component of the system.

A. Shape Features

To obtain the shape features we need to localize the face and certain points on it, either by manual human annotations or with the help of a face tracking system. In order to avoid any noise possibly introduced by automatic face tracking and to better observe the improvement provided by the proposed texture based features (explained in section II-B) we use manual annotations of 68 facial points for the tests presented in this paper.

The face is divided into three regions and only a certain group of the facial points are used corresponding to each region. The reason for doing this is that none of the action units causes a substantial change in the whole face or all of the 68 points defined, but only a specific portion. So, we can reduce the computational burden and noise caused by the feature extraction and selection processes. More precisely, we use 29 points and the texture contained inside and around for each of the upper face (AUs 1,2,4,5 and 7), middle face (AUs 6 and 9) and lower face (AUs 12,15,17,20,23,24,25 and 27) action units. The selected points for each type can be seen in Fig.1.

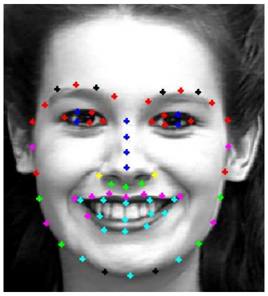


Fig. 1: Points used in feature extraction; Upper face points are shown in red, blue or yellow; Middle face points in green, blue, magenta or yellow; Lower face points in cyan, magenta or yellow; Black points are not taken into account for any AU

The shape features are then obtained using the initial frame (containing no expression) and peak expression frame (referred to as peak frame throughout the paper) of each video sequence containing an expression of N frames, similarly to [3] with the difference of using only 2 frames rather than the whole sequence. All of the shapes (68 points) were aligned to a single shape to exclude the effect of translation, rotation and scale. The first features obtained is the position change in horizontal and vertical directions of the 29 points defined, which is called set S. Thus, we compute for each point i in S

$$F_1(i) = x_{i,N} - x_{i,1} \tag{1}$$

$$F_2(i) = y_{i,N} - y_{i,1} \tag{2}$$

$$F_3(i) = \sqrt{(x_{i,N} - x_{i,1})^2 + (y_{i,N} - y_{i,1})^2} \tag{3}$$

where $x_{i,N}$ denotes the position in x coordinate of point i in frame number N, or the peak frame, and similarly $x_{i,1}$ that in the first, or neutral, frame.

Then, we also take as features the change in position of all points with respect to each other in the peak and initial frames, i.e.

$$\begin{aligned} F_4(i,j) &= \sqrt{(x_{i,N} - x_{j,N})^2 + (y_{i,N} - y_{j,N})^2} - && (4) \\ & \quad \sqrt{(x_{i,1} - x_{j,1})^2 + (y_{i,1} - y_{j,1})^2} && (5) \\ F_5(i,j) &= atan\frac{|y_{i,N} - y_{j,N}|}{|x_{i,N} - x_{j,N}|} - atan\frac{|y_{i,1} - y_{j,1}|}{|x_{i,1} - x_{j,1}|} && (6) \\ & && (7) \end{aligned}$$

for all points $i \neq j$ in S. Obtaining in the end the feature set $F_s = [F_1, F_2, F_3, F_4, F_5]$ of 899 shape features.

B. Texture Features

The texture related features that we propose to use, to explain simply, is the difference, between initial and peak frames, of the histograms that we obtain from overlapping windows of various sizes on the LBP transform that is applied on three images obtained by three different filters. These filters are the bilateral filter, opening by reconstruction filter and black top-hat by reconstruction filter. We explain in the following subsections how each of them works and why they are relevant to our task, in addition to the LBP transform and the feature extraction procedure.

1) Bilateral Filter: The first preprocessing method we perform in order to eliminate irrelevant facial deformations or noise present in the image is the bilateral filter. The bilateral filter is a non-linear filter introduced by Tomasi et al. [8] and has been vastly used mainly for the purposes of image denoising and for creating special effects in photographs. Its main advantage compared to linear filters is that it smoothes an image while preserving the edges with the help of two different kernels called the domain and range filter. The equation of the bilateral filter is given as

$$\hat{I}(p_c) = w_c^{-1} \sum_{k \in Q} e^{-\frac{||p_c - p_k||^2}{2\sigma_d^2}} e^{-\frac{(I(p_c) - I(p_k))^2}{2\sigma_r^2}} I(p_c) \tag{8}$$

where Q is the particular neighborhood taken around the pixel located at p_c and I denotes the corresponding gray-level intensity. The normalization factor w_c is simply the summation of the weights over the neighborhood Q.

The first kernel in (8) is the simple Gaussian smoothing filter, called the domain filter in this case. The second one, called range filter, is where the non-linearity appears and it smoothes the image in the intensity domain. This means that, the neighboring pixels with intensity values close to the center pixel are assigned a smaller weight than the pixels that have a larger intensity difference. Thus, the areas which contain edges (high intensity changes) are less affected by the smoothing performed by the domain filter.

The bilateral filter is suitable for our case, since our main source of information is contained on the edges created by the facial actions, and we want to smooth out the regions that contain other irrelevant deformations. The main issue with bilateral filters is the choice of the 3 parameters σ_d,σ_r and the neighborhood size, which affect directly the amount of smoothing and edge preserving. No optimization of these parameters exists in the literature and the optimal parameters depend highly on the application, so, in this preliminary work, we choose empirically as parameters $\sigma_d = 3$, $\sigma_r = 50$ and a square neighborhood of size 11, which provides a reasonable smoothing. An example result of the bilateral filter and the LBP transform applied on it can be seen in Fig.2b. As expected the LBP transform results in smoother regions, so that the main patterns explaining the facial features are better viewed and, of course, identified.

2) Morphological Operations by Reconstruction: The second type of preprocessing that we use is based on mathematical morphology. Opening and closing are two of the most commonly used morphological operations. Morphological opening serves to identify or isolate structures (or connected components) that are brighter than their environment while morphological closing isolates and flattens image structures that are darker than their surroundings and that have a smaller support than the structuring element (SE) used for the consecutive dilation and erosion operations. Depending on the structuring element, the way that the image behaves under these filters thus provides information on structural features of the objects present in the image. They have been frequently used to obtain feature sets using varying sizes of structural elements in tasks like image classification and segmentation, especially in remote sensing applications[9].

Based on this ability of defining bright and dark structures in images, we adopt the idea of using the morphological filters as a preprocessing method applied before the LBP transform. The standard opening and closing operations, however, result in the deformation of important geometrical structures as well. To prevent this severe effect, a shape preserving method called morphological filtering by reconstruction was proposed [10], with the idea of avoiding deformation of structures larger than the structuring element.

Opening and closing by reconstruction are performed in two steps. In the case of opening, first a marker image I_M is obtained by applying erosion (represented by ϵ) on the original image I, using the structural element B.

$$I_E = \epsilon_B(I) \tag{9}$$

The second phase is iteratively performing a geodesic dilation starting with the marker image I_E until no further change in the image pixels is obtained. The geodesic dilation on an image is defined simply as the pixel-wise minimum ($\wedge$) of the elementary dilation (dilation with the smallest structuring element, represented as δ_1) on the image and a mask image, which is in our case the original image, I [9]. After n iterations we obtain the opening by reconstruction, I_{OR}, in the form

$$I_{OR} = \delta^n_{1,I}(I_E) = \delta_{1,I}(\delta_{1,I} \ldots (\delta_{1,I}(I_E))) \tag{10}$$

with

$$\delta_{1,I}(I_E) = \wedge\{\delta_1(I_E), I\} \tag{11}$$

and

$$\delta^{n+1}_{1,I}(I_E) = \delta^n_{1,I}(I_E) \tag{12}$$

Closing by reconstruction (I_{CR}) is obtained, similarly, by iteratively applying the geodesic erosion operation on the marker image obtained by dilating the original image with a structural element B, until the resulting image is identical to the one in the previous iteration. The geodesic erosion is defined as the pixel-wise maximum ($\vee$) of the elementary erosion of the marker image and the mask image, which is once again our original image I.

$$I_{CR} = \epsilon^n_{1,I}(I_D) = \epsilon_{1,I}(\epsilon_{1,I} \ldots (\epsilon_{1,I}(I_D))) \tag{13}$$

with

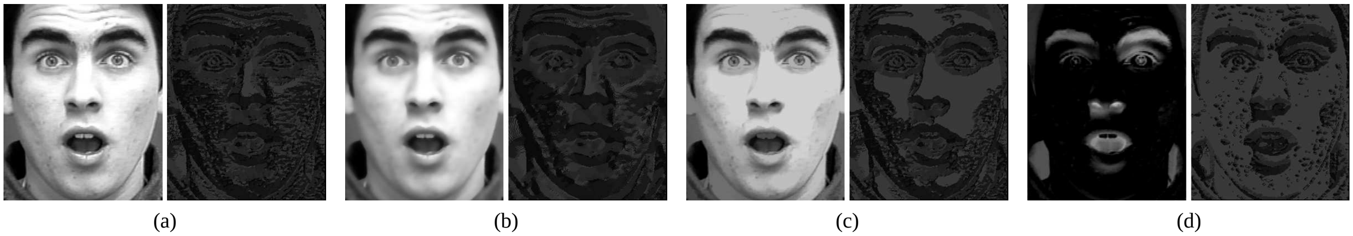
$$\epsilon_{1,I}(I_D) = \vee\{\epsilon_1(I_D), I\} \tag{14}$$

We use as our morphological preprocessing methods the opening by reconstruction and the black top-hat by reconstruction method. The black top-hat transform (also called the closing by top-hat or top-bottom transform) is the residual of a closing image when compared to the original image:

$$I_{BTR} = I_{CR} - I \tag{15}$$

Example results of the opening by reconstruction, black top-hat transform and the LBP transform applied on top can be seen in Fig.2c and 2d respectively. As we can see the opening performed serves to flatten the bright areas on the face, emphasizing the important intensity changes caused by the facial features, and to help the LBP transform obtain clearer structures. The black top-hat transform, on the other hand, identifies the dark regions on the face (such as the mouth opening and eyebrows) and therefore cause the LBP to have more significant boundaries around these regions. As the structuring element we use a disk shape of size 30 by 30 pixels, for images of size 640 by 490. All filter parameters were chosen based on visual observations for this initial work, but in future work we plan to optimize these parameters using cross-validation tests.

3) Feature Extraction by Uniform Local Binary Pattern Histogram Differences: Local binary Patterns (LBP) is an efficient gray-scale texture descriptor proposed by Ojala et al. [11] and has been used widely in various texture description and classification problems, including expression recognition and AU detection, along with its many variants. Its main advantage is that it is invariant to illumination changes since it is defined by the relationship of a pixel with its



(a) (b) (c) (d)

Fig. 2: Examples of the preprocessed images and their LBP transforms (on the right of each subfigure) (a) is the original image, (b) is the bilateral filtered image, (c) is the image after opening by reconstruction and (d) is the image after black top-hat by reconstruction

neighbors, thus can identify successfully the microstructures in an image. The basic LBP is defined for a pixel p_c as

$$LBP(p_c) = \sum_{k=0}^{P-1} l(I(p_k) - I(p_c)).2^k \tag{16}$$

where $I(p)$ denotes the intensity of a pixel p, and P is the total number of pixels in the chosen neighborhood of the center pixel p_c. The function l is a simple thresholding function in the form

$$l(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{otherwise} \end{cases} \tag{17}$$

In the end we obtain a binary pattern of P bits for each pixel. By varying this number P and the radius of the circular neighborhood one can obtain LBP at different resolutions. In this work we use the uniform LBP on a neighborhood of radius=1 and $P = 8$. Uniform LBP [12] is an extension of the standard LBP, where the binary patterns are grouped according to the number of 0/1 transitions that they contain, and the patterns containing more than 2 transitions (non-uniform patterns) are assigned the same identity, since it was shown that they occur much less frequently than the others, namely the 58 uniform patters. So, for each pixel in a region of interest we assign a value from 0 to 58, and obtain a 59 bin histogram for that region. Figure 2a shows the uniform LBP transformation of an example face image from the CK+ database [13].

In our experiments we scale each face region (upper, middle or lower as explained in Section IIA) in the initial and peak frames to a standard size of 240 to 120 pixels. Then we obtain the 59 bin uniform LBP histograms of 324 overlapping windows of different sizes, the smallest window size being 40 by 40 while the largest one is 240 by 120 containing the whole region of interest. Figure 3 shows an illustration of the windows with the smallest size along with the first two slid versions; the overlap size is $M_1 - 20$ by $M_2 - 20$ for each window of size M_1 by M_2. Most of the works to date using LBP histograms for action unit detection have used standard size non-overlapping windows. However, for each AU the most important information may be contained in windows of different sizes and positioned in various locations. For instance, for action unit 2 (outer brow raise) the large window containing both of the eye brows is intuitionally more important than the smaller window containing only the inner brows, while for action unit 1 (inner brow raise) it is not the case. Therefore, we prefer not to discard any of these overlapping regions, and let the feature selection step choose the most relevant ones.

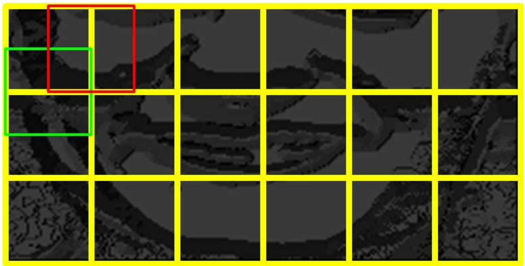


Fig. 3: Illustration showing the smallest window size used for LBP histogram extraction and the first two overlapping translated versions

Once we have obtained the histograms for each of the windows on each of the initial and peak frames, we compute the histogram variation between the two frames, the reason being, using the change in the LBP profiles rather than the profiles directly in the peak frame eliminates the variations due to identity and provides a stronger feature set [7]. Instead of the direct difference of 2 histograms and using every bin as separate features as done in [7], we use the χ^2 distance, D_{χ^2}, which is defined as

$$D_{\chi^2}(H_N, H_1) = \sum_{b \in B} \frac{(H_N(b) - H_1(b))^2}{(H_N(b) + H_1(b))/2} \tag{18}$$

where $H_N(b)$ denotes the value at bin b of the histogram for the Nth frame, and B denotes the set of all the bins. The texture features for the region of concern is thus these distance measures for each of the 324 windows.

Applying the LBP transform and obtaining these texture features explained, for all three of the preprocessed images (bilateral filter, opening by reconstruction, black top-hat by reconstruction) we have our final set of 972 texture related features. The three different filtering methods, combined with the local binary pattern transform, allow us to obtain an extended set of features explaining the facial structure and as presented in the next section provide a much better AU detection accuracy compared to the LBP used alone, both in combination with the shape features and by themselves.

C. Feature Selection and Classification

Once the full set of features (shape + texture) is obtained, we perform feature selection using the GentleBoost algorithm [14] to choose the most relevant features for each of the AUs. We therefore perform this process 15 times independently, for the action units 1,2,4,5,6,7,9,12,15,17,20,23,24,25 and 27. Feature selection is a crucial step in the AU detection process, since it discards the irrelevant and redundant features which constitutes a huge portion of the total number of features extracted, due to the large number of LBP windows and inter-point relations we use for building our features set. For each action unit 200 features are extracted in total as result of the GentleBoost, then the optimal number of features is chosen by performing leave-one-subject-out tests (explained in detail in Section III) with 30,50,100,150 and 200 features for each AU separately.

For the detection of action units using these selected features, we train 15 Support Vector Machines (SVM), once again for each AU. The SVM are binary, the classes being if the specific AU is present in the image sequence or not. As kernels we use Gaussian Radial Basis Functions (RBF), and optimize the classifier parameters σ and C using a 5-fold cross validation on the dataset. For the SVM classification we use the publicly available LibSVM library[15]. The cross-validation tests and parameter optimization are explained in more detail in the results section.

III. EXPERIMENTAL RESULTS

For all the experiments that we performed we have used the Extended Cohn-Kanade (CK+) database [13], which consists of a total of 593 image sequences of 123 different subjects posing in various facial expressions and contains different numbers of examples of many action units. The action units present on the peak frame of each sequence were identified by human coders for each sequence. We have applied our methods to detect 15 action units which have a reasonable number of occurrences in the database. We take, for each AU, as positive examples all the sequences that it is present in the peak frame, regardless of the intensity of the action.

For each of the tests presented, we have performed a leave-one-subject-out (LOO) cross-validation; meaning, all sequences of a specific subject were excluded in the set used to train the classifier, then the classifier was tested on the excluded sequences and the overall accuracy was calculated by adding the number of correctly classified sequences for each subject. The best parameters set $\{\sigma, C\}$ of the SVM (corresponding to the highest classification rate) were chosen out of a possible 25, using a 5-fold cross validation on the training set for each subject. The LOO tests were performed for each AU using 30,50,100,150 and 200 features and the one giving the highest overall accuracy was chosen as the final result.

We group the results we obtained in two parts: The first one is the AU detection performance using only texture features in the feature selection and classification, and compares the two results obtained by the preprocessing methods, explained in Section IIB, applied before the LBP transform and by the LBP transform applied directly on the original image. The second part presents the detection results obtained by using these texture features in conjunction with the geometric features detailed in Section IIA, and compares these results to other methods in the literature that have reported results on the same database.

A. Experiments with only texture features

First, we train our feature selector and classifiers using only the texture features, not including yet the geometric features, in order to observe the advantage of applying the preprocessing methods proposed over using LBP transform directly on the image by itself. Table I presents the number of features used, overall accuracy and area under the receiver operator characteristics (ROC) curves, which are presented in Fig.4, for each of the 15 action units and for both methods. The overall accuracy (OA) stands for the correct classification rate for both the positive and negative examples for each AU.

We can see from these results the significant increase in accuracy when we use the extended set of texture features; that is, with the preprocessing applied. For all AUs we obtain a higher accuracy and AUC with the feature extraction method using the filters, resulting in an average increase of 2.34% in the OA, 4.57% in the AUC, which is statistically more meaningful than the OA due to the unbalanced number of positive and negative examples. The number of features giving the highest accuracy in each case is particularly interesting, since for certain AUs this number is higher for the method using only LBP, although the total number of features before feature selection is only one third of the other method (324 vs. 972). This fact serves to show us that the increase in accuracy is not at all dependent on the number of features extracted but rather on their ability to describe the facial actions.

These tests show not only the advantage of the preprocessing methods proposed, but also the potential of the system when it is completely automated, which is the next planned step. The texture features are mostly independent from the facial point annotations, for which we use manual annotations at this step, except for obtaining the relevant region part of the face. This can be easily and efficiently performed using existing face detection methods in the literature and we see, as explained in the following section and presented in Table II that we achieve accuracy measures competitive with other state-of-the-art methods even using only texture features.

B. Experiments with shape and texture features combined

The second group of experiments we perform is using the shape features explained in Section IIA in combination with the texture features explained in Section IIB. Once

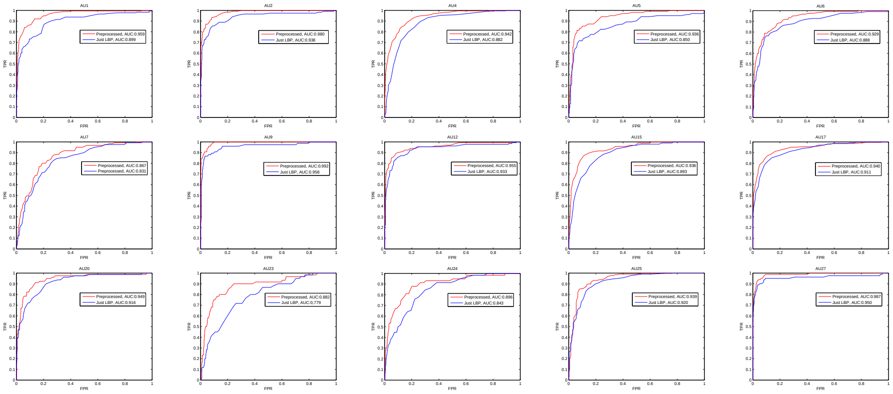


Fig. 4: Receiver Operator Characteristics curves for each of the Action Units included in the experiments. Red curves are the ones obtained using Preprocessing and LBP texture features, while blue curves are the ones obtained using only LBP texture features

TABLE I: AU Detection Results for the preprocessing + LBP texture features (Pre+LBP) and for only LBP texture features (LBP). NP: Number of positive examples for the AU in the database, nFts: Number of features used, OA(%): Percentage overall accuracy, AUC(%): Area under ROC curve

AU	NP	nFts		OA(%)		AUC(%)	
		Pre+LBP	LBP	Pre+LBP	LBP	Pre+LBP	LBP
1	177	200	150	**90.89**	86.17	**95.90**	88.87
2	117	200	150	**94.09**	93.25	**97.69**	92.98
4	194	100	150	**87.86**	82.13	**94.13**	88.18
5	102	30	100	**91.91**	89.04	**93.58**	87.14
6	123	100	150	**89.04**	86.85	**92.87**	88.51
7	121	100	150	**83.64**	83.61	**86.67**	83.07
9	75	50	150	**97.47**	95.45	**99.18**	95.82
12	131	150	50	**93.42**	90.89	**95.41**	93.30
15	94	50	150	**92.24**	88.02	**93.64**	89.36
17	202	150	50	**89.54**	86.51	**94.03**	91.13
20	79	100	100	**92.58**	92.07	**94.88**	91.57
23	60	100	150	**92.24**	89.38	**88.13**	77.98
24	58	50	150	**92.58**	91.39	**89.62**	84.30
25	324	150	100	**88.03**	85.83	**93.72**	92.48
27	81	100	150	**96.12**	95.95	**98.70**	95.02
Avg.				**91.44**	89.10	**93.88**	89.31

again we conduct the experiments using the LBP on top of 3 preprocessing methods, and using LBP directly on the image separately. In the first case the feature selection algorithm is fed 1871 features in total, while in the second this number is 1231. In this preliminary study aiming to test the efficiency of the proposed texture features we use only manual annotations of the facial points. Due to the high accuracy of these features and the ratio of the shape vs. texture features, the feature selection tends to select shape features more frequently in the LBP features without preprocessing case, as expected. Therefore, the difference in accuracies obtained by the two different methods is less significant than that presented in Section IIIA. With the preprocessed features we obtain 94.74% overall accuracy and 96.97% AUC, while with only the LBP features we obtain 94.13% accuracy and 96.01% AUC as average over the 15 AUs tested.

The preprocessed features achieve higher accuracy and AUC for 12 AUs, the exceptions being AU 23 and 24 for only the overall accuracy, which is rather meaningless since they have very few positive examples, and AU25 (jaw drop) for both accuracy and AUC, which has proven by the performance difference between using shape+texture features and only texture features (shown in Table II), to be very dependent on the features provided by the geometry of the facial points rather than the texture. Comparing these two performances (shape+texture vs. texture) we see that while shape features bring about a higher accuracy in all AUs, for some of them this change is more substantial, like AU1 (inner brow raise) in addition to AU25. This tells us that for these AUs change of location of facial points contains more important information than the change in texture contained in or around. It makes complete sense in the case of AU1 and AU25, for example, where we do not see a significant texture variation on the area related to these actions but an obvious position change of certain facial points.

We also compare our results to 3 different methods that have reported results on the same database. The first one is the method by Senechal et al. [7] in which they use as features the histogram differences of Local Gabor Binary Patterns(LGBP) in non-overlapping fixed size windows, and build a special kernel using this difference for the classifier. Since separate AU performances were not reported, and the lower AUs are not the same ones tested in this work we can only compare the mean upper AU detection performance. The best results that they achieve is with the special kernel which is 97.3% AUC, while for us this measure is 96.8%. With, the Gaussian RBF kernel, however, they achieve 96.2%, from which we can deduct that with a much lower number of features selected efficiently, higher performances can be achieved.

The comparison with the other two methods can be seen in Table II for the 13 common AUs that were tested in all three papers. The first method [3] proposes using as features only the position change of facial points throughout the whole sequence and does not report the AUC measure so we compare the F1 measure instead, noting that we tune our parameters to give the highest classification accuracy and

TABLE II: AU Detection Results comparison using our method with shape + texture features (SHTXT), our method with texture features only (TXT), the method proposed by Valstar & Pantic [3](Valstar) and the method proposed by Bartlett et al. [2] (Bartlett). OA: Overall accuracy, F1: F1 measure, AUC: Area under ROC curve

AU	OA				F1			AUC		
	SHTXT	TXT	Valstar	Bartlett	SHTXT	TXT	Valstar	SHTXT	TXT	Bartlett
1	0.965	0.909	0.918	0.92	0.938	0.841	0.826	0.983	0.959	0.95
2	0.976	0.941	0.939	0.88	0.939	0.836	0.833	0.991	0.977	0.92
4	0.911	0.879	0.870	0.89	0.862	0.809	0.630	0.968	0.942	0.91
5	0.944	0.919	0.904	0.92	0.829	0.745	0.596	0.976	0.936	0.96
6	0.911	0.890	0.930	0.93	0.778	0.716	0.811	0.946	0.929	0.96
7	0.882	0.836	0.870	0.88	0.688	0.531	0.290	0.917	0.867	0.95
9	0.992	0.975	0.928	1	0.966	0.895	0.573	0.998	0.992	1
12	0.944	0.934	0.930	0.95	0.865	0.838	0.836	0.974	0.954	0.98
15	0.953	0.922	0.969	0.85	0.839	0.726	0.361	0.956	0.936	0.91
20	0.963	0.926	0.908	0.92	0.849	0.690	0.517	0.973	0.949	0.84
24	0.946	0.926	0.935	0.92	0.682	0.511	0.497	0.945	0.896	0.88
25	0.959	0.880	0.851	0.89	0.963	0.889	0.748	0.984	0.937	0.93
27	0.985	0.961	0.964	0.99	0.945	0.855	0.854	0.996	0.987	1
Avg.	0.949	0.915	0.916	0.909	0.857	0.760	0.638	0.969	0.943	0.926

not the highest F1. The second method [2] uses only Gabor features with an Adaboost classifier. We achieve in average, and for most of the action units, superior performance compared to the 2 methods, both when we use shape and texture features together and when we use only the texture features. Once again, the shape features we use depend highly on the accuracy of the facial points, for which we have only used human annotations at this stage, but the promising accuracy measures obtained for both types of features already show the strength of the proposed features in detecting action units.

IV. CONCLUSIONS

In this paper we have presented a simple, novel and efficient method for identifying features for Action Unit detection in videos that is based on Local Binary Patterns applied separately to images processed by three different filtering methods, namely the bilateral filter, opening by reconstruction and black top-hat by reconstruction. The results obtained show that this method provides a significant increase in the accuracy measures for all 15 action units tested compared to using LBP by itself.

We have also used the extracted texture related features along with certain transient geometric features, and demonstrated that we achieve performances superior to existing approaches tested on the same database. Testing the system with a facial point tracking system for complete automation is the next step in the process. Our experiments using only texture features, which are mainly independent of the tracked points, result in very high performances already, proving the strength of the features proposed in detecting facial actions.

V. ACKNOWLEDGMENTS

This work has been supported by the Swiss National Science Foundation through the National Centre for Competence in Research(NCCR) on Interactive Multimodal Information Management(IM2).

REFERENCES

[1] Ekman, P.; Friesen, W.; ”The Facial Action Coding System: A Technique for the Measurement of Facial Movement”, San Francisco CA: Consulting Psychologists Press, Inc., 1978.

[2] Bartlett, M.; Littlewort, G.; Frank, M.; Fasel, I.; Movellan, J.;”Automatic system for measuring facial expression in video”, in *Journal of Multimedia*, vol:1, no.6, pp.22-35, September 2006.

[3] Valstar, M.F.; Pantic, M.; ”Fully Automatic Recognition of the Temporal Phases of Facial Actions,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* , vol.42, no.1, pp.28-43, Feb. 2012.

[4] Valstar, M.F.; Mehu, M.; Bihan Jiang; Pantic, M. ; Scherer, K.; ”Meta-Analysis of the First Facial Expression Recognition Challenge”, in *Systems, Man, Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no.4, pp.966-979, August 2012.

[5] Senechal, T.; Rapp, V.; Salam, H.; Seguier, R.; Bailly, K.; Prevost, L.; ”Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units,” *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* , vol., no., pp.860-865, 21-25 March 2011.

[6] Shan, C.; Gong, C.; McOwan, P.W.; ”Facial expression recognition based on Local Binary Patterns: A comprehensive study,” *Image and Vision Computing, Journal of*, vol.27, no.6, pp.803-816, May 2009.

[7] Senechal, T.; Bailly, K.; Prevost, L.; ”Automatic Facial Action Detection Using Histogram Variation Between Emotional States,” *Pattern Recognition (ICPR), 2010 20th International Conference on* , pp.3752-3755, 23-26 Aug. 2010.

[8] Tomasi, C.; Manduchi, R.; , ”Bilateral filtering for gray and color images,” *Computer Vision, 1998. Sixth International Conference on* , pp.839-846, 4-7 Jan 1998.

[9] Dalla Mura, M.; Benediktsson, J.A.; Chanussot, J.; Bruzzone, L; ”The Evolution of the Morphological Profile: from Panchromatic to Hyperspectral Images,” in *Optical Remote Sensing, Augmented Vision and Reality*,vol.3, Berlin, Germany: Springer-Verlag Berlin Heidelberg, 2011, pp.123-146.

[10] Crespoa, J.; Serra, J.;, Schafer, R.; ”Theoretical Aspects of Morphological Filters by Reconstruction,” *Signal Process.*, vol. 47, no. 2, pp. 201-225, Nov. 1995.

[11] Ojala, T.; Pietikainen, M.; Harwood, D.; ”A comparative study of texture measures with classification based on featured distributions”, *Pattern Recognition*, vol.29, no.1, pp.51-59, 1996.

[12] Ojala, T.; Pietikainen, M.; Maenpaa, T.; ”Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971-987, 2002.

[13] Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I.; ”The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,” *Computer Vision and Pattern Recognition Workshops (CVPRW) 2010 IEEE Computer Society Conference on* , pp.94-101, 13-18 June 2010.

[14] Friedman, J.; Hastie, T.; Tibrishirani, R.; ”Additive Logistic Regression: a Statistical View of Boosting” in *Annals of Statistics*, vol:28, no.2, pp.337-407, 2000.

[15] Chang, C.C.; Lin, C.J.; ”Libsvm: A library for support vector machines,” *Intelligent Systems and Technology, ACM Transactions on*,vol. 2 no.3 pp.27:1-27:27,May 2011.