A STUDY FOR THE SELF SIMILARITY SMILE DETECTION

D. Freire, L. Antón, M. Castrillón.

SIANI, Universidad de Las Palmas de Gran Canaria, Spain
dfreire@iusiani.ulpgc.es,
lanton@iusiani.ulpgc.es,mcastrillon@iusiani.ulpgc.es

Abstract. Facial expression recognition has been the subject of much research in the last years within the Computer Vision community. The detection of smiles, however, has received less attention. Its distinctive configuration may pose less problem than other, at times subtle, expressions. On the other hand, smiles can still be very useful as a measure of happiness, enjoyment or even approval. Geometrical or local-based detection approaches like the use of lip edges may not be robust enough and thus researchers have focused on applying machine learning to appearance-based and self-similarity descriptors. This work makes an extensive experimental study of smile detection testing the Local Binary Patterns (LBP) combined with self similarity (LAC) as main descriptors of the image, along with the powerful Support Vector Machines classifier. Results show that error rates can be acceptable and the self similarity approach for the detection of smiles is suitable for real-time interaction, although there is still room for improvement.

1 Introduction

It is now known that emotions play a significant role in human decision making processes [14]. The ability to show and interpret emotions is therefore also important for human-machine interaction. In this context face analysis is currently a topic of intensive research within the Computer Vision community. Facial expression recognition research has studied geometry-based features [3], appearance [1] and hybrid approaches [7], see [11] for a survey. Commercial products that are able to perform expression recognition in real time are currently available. Potential applications include evaluation of behavior, human-robot interaction, intelligent tutoring systems, perceptual user interfaces, etc.

Some facial expressions can be very subtle and difficult to recognize even between humans. Besides, in human-computer interaction the range of expressions displayed is typically reduced. In front of a computer, for example, subjects rarely display accentuated surprise or anger expressions as he/she would display when interacting with another human subject.

The human smile is a distinct facial configuration that could be recognized by a computer with greater precision and robustness. Besides, it is a significantly useful facial expression, as it allows to sense happiness or enjoyment and even approval (and also the lack of them) [8]. As opposed to facial expression recognition, smile detection research has produced less literature. Lip edge features and a perceptron were used in

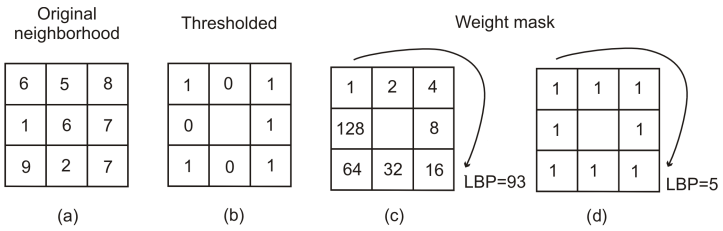

Fig. 1. The basic version of the Local Binary Pattern computation (c) and the Simplified LBP codification (d).

[10]. The lip zone is obviously the most important, since human smiles involve mainly the Zygomatic muscle pair, which raises the mouth ends. Edge features alone, however, may be insufficient.

We present an image descriptor based on self-similarities which is able to capture the general structure of an image. Computed descriptors are similar for images with the same layout, even if textures and colors are different. Similarly to [16], images are partitioned into smaller cells which, conveniently compared with a patch located at the image center, yield a vector of values that describes local aspect correspondences (LAC)

This paper makes an extensive experimental study of the smile detection problem, being organized as follows. Section 2 describes the codification algorithms used for the experiments. The different classification approaches used in the study are briefly presented in Section 3. The experimental results and conclusions are described in Sections 4 and 5 respectively.

2 Representation

The Local Binary Pattern (LBP) is an image descriptor commonly used for classification and retrieval. Introduced by Ojala et al. [13] for texture classification, they are characterized by invariance to monotonic changes in illumination and low processing cost. Given a pixel, the LBP operator thresholds the circular neighborhood within a distance by the pixel gray value, and labels the center pixel considering the result as a binary pattern. The basic version considers the pixel as the center of a 3×3 window and builds the binary pattern based on the eight neighbors of the center pixel, as shown in Figure 1-c. However, the LBP definition can be easily extended to any radius, R, considering P neighbors [13]:

Rotation invariance is achieved in the LBP based representation considering the local binary pattern as circular.

More recently LBPs have been used to describe facial appearance. Once the LBP image is obtained, most authors apply a histogram based representation approach [15]. However, as pointed out by some recent works, the histogram based representation loses relative location information [15, 17], thus LBP can also be used as a preprocessing method. Using LBP as preprocessing method, has the effect of emphasizing edges and

noise. To reduce the noise influence, Qian Tao et al. [17] proposed recently a modification in the basic version of the local binary pattern computation. Instead of weighting the neighbors differently, their weights are all the same, obtaining the so called Simplified LBPs, see Figure 1-d. Their approach has shown some benefits applied to facial verification, due to the fact that by simplifying the weights, the image becomes more robust to illumination changes, having a maximum of nine different values per pixel. The total number of local patterns are largely reduced so the image has a more constrained value domain.

In the experiments described at Section 4, both approaches will be adopted, i.e. using the histogram based approach, but also using Uniform LBP and Simplified LBP as a preprocessing step.

Raw face images are highly dimensional. A classical technique applied for face representation to avoid the consequent processing overload problem is Principal Components Analysis (PCA) decomposition [12]. PCA decomposition is a method that reduces data dimensionality, without a significant loss of information, by performing a covariance analysis between factors. As such, it is suitable for highly dimensional data sets, such as face images. A normalized image of the target object, i.e. a face, is projected in the PCA space. The appearance of the different individuals is then represented in a space of lower dimensionality by means of a number of those resulting coefficients, $\mathrm{v}_i[18]$.

We also present an image descriptor based on self-similarities which is able to capture the general structure of an image. Computed descriptors are similar for images with the same layout, even if textures and colors are different, similarly to [16]. Images are partitioned into smaller cells which, conveniently compared with a patch located at the image center, yield a vector of comparison results that describes local aspect correspondences (LAC).

A LAC descriptor is computed from a square shaped image subdivided into $n \times n$ cells, where each cell corresponds to an $m \times m$ pixels image patch. The number of cells and their pixel size have effect on how much an image structure is generalized. A low number of cells will not capture many structural details, while too many small cells will produce a too detailed descriptor. The present approach will consider overlapping cells, which may be required to capture subtle structural details.

Once an image is partitioned, an $m \times m$ patch located in the exact image center (which does not have to correspond to a cell in the image partition) is compared with all partition cells. In order to achieve greater generalization, image patches are compared computing the Sum of Squared Differences (SSD) between pixel values (or the Sum of Absolute Differences (SAD), which is computationally less expensive). Each cell-center comparison is consecutively stored in a $m \times m$ dimensions LAC descriptor vector.

Such description overcomes color, contrast and textures. Images are described in terms of their general structure, similarly to [16]. An image showing a white upper half and a black lower half will produce exactly the same descriptor as an image showing a black upper half and a white lower half. Local aspect correspondences are exactly the same: the upper half is different from the lower half. Rotations, however, are not considered.

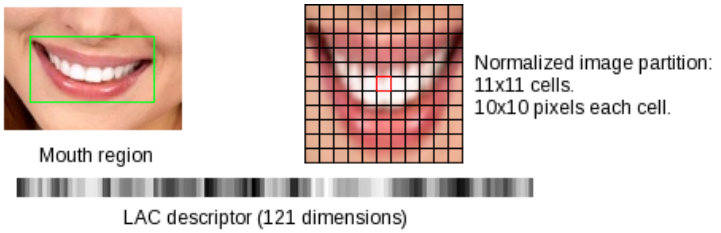

Fig. 2. LAC Descriptor example using a 11x11 partition. The barcode-like vector represents all comparisons between each cell and the central patch.

LAC descriptors are specially useful to describe points defined by a scale salient point detector (like DoG or SURF [2]). In the present work, however, they are applied to classify mouths found by a face detector [5] into smiling or non-smiling gestures. Smiling mouths look similar no matter the skin color or the presence of facial hair. This generality can be registered by a self-similarity descriptor like LAC.

However, images containing smiling mouths require local brightness to be preserved: teeth are always brighter than surrounding skin and that must be captured by the descriptor. Thus, instead of using SSD, patches are compared using Sum of Differences. Otherwise, a closed mouth would produce the same descriptor as a smiling mouth: lips are surrounded by differently colored skin, exactly as teeth are surrounded by differently colored lips. Figure 2 shows an example with an 11×11 cell partition, each cell sized 10×10 pixels. The LAC descriptor is shown as a barcode for representation purposes.

Thus, given an input image (i.e. a scale salient point or a known region like detected mouths), a number of cells n and their pixel size m, LAC is computed as follows:

1. The image is resized to a template sized $(n \times m) \times (n \times m)$ pixels.
2. The template is partitioned into $n \times n$ cells, each of them sized $m \times m$ pixels.
3. A central patch sized $m \times m$ pixels is captured from the center of the template image.
4. The central patch is compared with each template cell, and each result is consecutively stored in the $n \times n$ LAC descriptor vector.

In order to tell wether two images have a similar structure, their corresponding LAC descriptors can be compared computing SAD between both vectors. However, given that the present work aims at classifying mouth images in two categories, a Support Vector Machine approach is used.

3 Classification

Support Vector Machine (SVM) [4] is a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers. A property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as maximum

margin classifiers. LIBSVM [6] has been the library employed in the experiment described below.

4 Experiments

The dataset of images used for the experimental setup is separated into two classes: smiling and not smiling. The previous classification has been performed by humans who labeled each normalized image of 59×65 pixels. The first set contains 2421 images of different smiling faces, while the second set contains 3360 non smiling faces.

As briefly mentioned above, the experimental setup considers one possibility as input: the mouth.

The input image will be a grayscale image, and for representation purposes we have used the following approaches for the tests:

- A PCA space obtained from the original gray images of 59×65 pixels.
- A PCA space computed on both the resulting images after preprocessing the original images using LBP. Two different approaches, i.e. simplified LBP (SLBP) and uniform LBP (ULBP), have been used.
- A concatenation of histograms based on the gray image or the resulting LBP image (both approaches simplified and uniform were used).
- A concatenation of the image values based on the gray images or the resulting LBP image (again both approaches simplified and uniform were used).
- LAC descriptor obtained from the original gray images of 59×65 pixels.
- LAC descriptor computed on both the resulting images after preprocessing the original images using LBP. Two different approaches, i.e. simplified LBP (SLBP) and uniform LBP (ULBP), have been used.

Similar experimental conditions have been used for every approach considered in this setup. The test sets are built randomly, having an identical number of images for both classes. Results presented correspond to the percentage of wrong classified samples of all test samples.

Average results presented in this paper are achieved for each configuration after ten random selections with 50% of samples for training and 50% for testing. Therefore, 2000 images, 1000 of each class, and 2000 images for the test, 1000 of each class, have been used for testing purposes.

As it can be seen in Figure 3, best results in almost every situation are achieved with no preprocessing at all, directly using grayscale images. None of the LBP based representations outperforms that approach. However, even if the Uniform LBP approach evidences a larger improvement when normalized histograms are used, the Simplified LBP approach reported better results than Uniform LBP in any other situation. As already stated in [17] this preprocessing provides benefits in the context of facial analysis.

When the Normalized Histogram based representation is used, the Uniform LBP error rate is the lowest. This approach seems to model properly the smile texture even when the histogram is losing the relative location information. However, this feature is quite similar for the Simplified LBP approach, its histogram loses information but achieved rates are similar. The grayscale image test achieved its highest error rate in

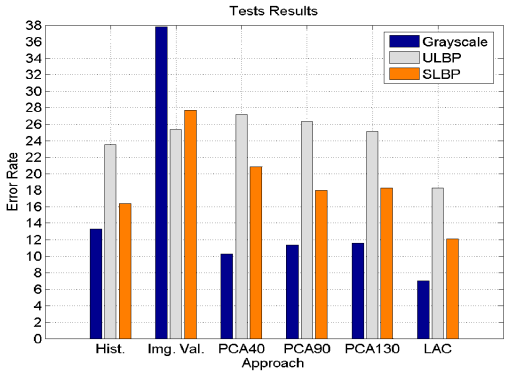

Fig. 3. Mouth processing results with different approaches using SVM for classification. Six different methods were applied to each preprocessing method: Histogram, Image Values, Principal Components Analysis and Local Aspect Correspondence respectively. It is important to mention that the number next to PCA refers to the dimension of the representation space, i.e. it indicates the number of eigenvectors used for projecting the face image.

this case, higher than the Uniform LBP and Simplified LBP approaches, which means that the Grayscale approach is very sensitive to the relative location of the information.

When the Normalized Image Values vector based representation is used, the grayscale image test achieved the lowest error rate. On the other hand, Uniform LBP test achieved the highest error rate in this case.

For the PCA approach, error-rate behavior is quite similar to the behaviour obtained previously with the rates of the image values test. Again grayscale image test achieved the lowest error rates. PCA deserves an additional observation, not always the increasing of the space dimension for PCA reports better results.

For the LAC approach, error-rate behavior is also quite similar to behaviour obtained previously with PCA and Image Value tests rates. Again, the lowest error rates were achieved by the grayscale image test. For LAC, it is important to mention that overlap is considered between cells. Firstly, several tests without overlapping were made in order to find optimun LAC parameters (number of cells and cells' size yielding the lowest error rate). For smile detection it was found that 10×10 cells of 3×3 pixels performed best thanks to the closeness between the size of the extracted LAC patch (30×30 pixels) and the original size of the mouth capture (20×12 pixels). It is also shown that worst results are achieved for configurations less than 10×10 cells because of the loss of information due to resizing in the Normalization step. Beyond that number of cells and for bigger sizes, behaviour is irregular due to the fact that information extracted is not reliable because of the false information introduced when the mouth is resized to fit the LAC patch. When images are upsampled, redundant and useless information is created. Unfortunately, when overlap was introduced, the achieved error rates were higher than without overlapping. Used images were too small for overlapping regions to be significant.

Something that should be mentioned is that, in terms of error rates, Simplified LBP behaves as an intermediate approach between Grayscale and Uniform LBP. That is, Simplified LBP has achieved better rates for normalized Histogram test than Grayscale's approach and worse than Uniform LBP's. Also Simplified LBP has got better rates for image values and PCA tests than Uniform LBP's approach and worse than the Grayscale's.

Unlike the study stated in [9] where whole face was considered for smile detection, for the SVM setting already explained, the strategical block of mouth is translated into a reduction of dimensions. Improvement is due to this fact. Of course, it should be mentioned that PCA reduce dimensions too, that is the reason why PCA tests achieved better results than the Image Values test for grayscale images.

Between approaches used in the tests, the difference of rates could not come from the domain value. Every input to SVM is previously normalized within the range [0,1].

Normalized Histograms deserves an additional observation. For each representation approach, a normalized histogram is built for the selected area: mouth. We can appreciate that, for the results in this case, there is a remarkable improvement of Uniform LBP above Simplified LBP.

5 Conclusions

This paper described a smile detection using different LBP approaches, as well as grayscale image representation, combined with SVM. It has been shown the potentiality of the LAC based representation for smile verification. The LAC based representation presented in this paper outperforms other approaches with an improve over a 5% for each preprocessing method. Overlap does not perform better due to the small size of the mouth area.

Uniform LBP does not respond to a statistical spatial patterns locality. This means, that there is no gradual change between adjacent blocks preprocessed with Uniform LBP. Depending on the value of a pixel inside one of the blocks, codification between two adjacent pixels can be, for example, from pattern 2 to pattern 9. Translated to the space domain of SVM, this means that dimensions can be too far away.

Simplified LBP keeps the statistical spatial patterns locality. There is a gradual change between adjacent preprocessed pixels. Translated to the SVM's space domain, this gradual change means that similar points are closer in this space.

The main reason to get worse results using a histogram of Simplified LBP, is that there is a loss of information related to location. That is not important for Uniform LBP because, as we said before, Uniform LBP looks for texture and the histogram window gives it the chance to show this fact. Our future line is focus on the potentiality of the LAC descriptor for generic applications such as image retrieval. In this paper we have developed a static smile clasiffier achieving, in some cases, a 93% of success rate. Due to this sucess rate, smile detection in video streams, where temporal coherence is implicit, will be studied in short term, as a cue to get the ability to recognize the dynamics of the smile expression.

References

1. M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *In Proceedings of the Seventh IEEE International Conference on Automatic Face and Gesture Recognition*, 2006.
2. H. Bay and T. Tuytelaars. Surf: Speeded up robust features. In *Proceedings of the Ninth European Conference on Computer Vision*, May 2006.
3. F. Bourel, C. Chibelushi, and A. Low. Robust facial expression recognition using a state-based model of spatially-localised facial dynamics. In *In Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.
4. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.
5. M. Castrillón Santana, O. Déniz Suárez, M. Hernández Tejera, and C. Guerra Artal. ENCARA2: Real-time detection of multiple faces at different resolutions in video streams. *Journal of Visual Communication and Image Representation*, pages 130–140, April 2007.
6. C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/ cjlin/ libsvm`.
7. D. Datcu and L. Rothkrantz. Automatic recognition of facial expressions using bayesian belief networks. In *In Proceedings of IEEE Systems, Man and Cybernetics*, 2004.
8. P. Ekman and W. Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–252, 1982.
9. D. Freire, M. Castrillon, and O. Deniz. Smile detection using local binary patterns and support vector machines. In *In Proceedings of the Fourth International Joint Conference on Computer Vision and Computer Graphics Theory and Applications*, 2009.
10. A. Ito, X. Wang, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Procs. of the 2005 IEEE Int. Conf. on Cyberworlds (CW'05)*, 2005.
11. M. Khan, M. Ingleby, and R. Ward. Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations. *ACM Trans. Auton. Adapt. Syst.*, 1(1):91–113, 2006.
12. Y. Kirby and L. Sirovich. Application of the Karhunen-Loéve procedure for the characterization of human faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1), July 1990.
13. T. Ojala, M. Pietikinen, and T. Menp. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.
14. R. W. Picard. *Affective Computing*. MIT Press, 1997.
15. Y. R. Sébastien Marcel and G. Heusch. On the recent use of local binary patterns for face authentication. *International Journal of Image and Video Preprocessing, Special Issue on Facial Image Processing*, 2007.
16. E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition 2007 (CVPR'07)*, June 2007.
17. Q. Tao and R. Veldhuis. Illumination normalization based on simplified local binary patterns for a face verification system. In *Proc. of the Biometrics Symposium*, pages 1–6, 2007.
18. M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.