Improving faces/non-faces discrimination in video sequences by using a local spatio-temporal representation

Yoanna Martinez-Diaz, Heydi Mendez-Vazquez, Noslen Hernandez, Edel Garcia-Reyes

Improving Faces/Non-Faces Discrimination in Video Sequences by Using a Local Spatio-Temporal Representation

Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Noslen Hernández, Edel García-Reyes
Advanced Technologies Application Center
7a #21406 b/ 214 and 216, P.C. 12200,
Siboney, Playa, Havana, Cuba
{ymartinez,hmendez,nhernandez,egarcia}@cenatav.co.cu

Abstract

In the last years spatio-temporal representations have shown to be successful for the analysis of video sequences in applications such as event detection, action and face recognition in videos. In this paper, we propose the use of local spatio-temporal features for the faces/non-faces classification stage, in the process of face detection in videos. Specifically, the extension of the Local Binary Patterns operator to the spatio-temporal domain is evaluated and compared with other schemes based on the same operator without considering the temporal information. The obtained results in the very challenged YouTube Faces database show that combining local appearance with motion can help to discriminate between faces and non-faces in the context of video applications.

1. Introduction

Face detection is a very important task for many video applications such as visual surveillance and human-computer interaction systems [19]. The more precise this step, the more accurate any further processing. Usually, it is also required as the first step in the process of video face recognition. However, most of face recognition methods assume that faces in a video have been already detected and localized. Therefore, in order to build fully automated video face recognition systems, robust and efficient face detection algorithms are needed.

Different approaches have been proposed to obtain the location of faces at every frame of a video. The simplest method is to detect the faces frame-by-frame [3]. However, this strategy can be time consuming because it needs to scan each frame at various scales and locations, in order to find existing faces. Besides, this approach does not exploit the temporal information available in a video. A second method is to detect the face in the first frame and then use a face tracking algorithm [18]. The main problems of this technique are to recover the tracking once it is lost and to identify the entry of new faces on the scene [7]. Another approach is to apply existing methods to detect moving objects in videos [12] to the case of faces. Among this kind of methods, the most successful ones are those that make use of the temporal information by modeling the background and finding motion patterns [11]. Nevertheless, most of them only give good results when the background changes slowly or it is a static scene [12], not performing well on natural scenes.

Although there are a number of algorithms available to find face regions in a video sequence, this paper is concerned to face detection as an input to a video face recognition system. It has been shown in different studies the benefits of using spatio-temporal information for video face recognition [1, 15]. Most of methods developed for this purpose need to have the temporal order of a face in a video sequence. Hence, we propose to use spatio-temporal representations for face detection in video; which could help to avoid some of the aforementioned problems. Different from frame-by-frame detection methods, using a spatio-temporal descriptor will allows us to detect faces in a set of frames at the same time. Besides, although the process is repeated every certain number of frames, it is not necessary to scan every frame of the sequence looking for faces; so, the processing time could be reduced. Moreover, using the same feature space for both face detection and face recognition, could bring some advantages [5].

Despite the fact that spatio-temporal representations have been widely used in many applications such as human action recognition [8], event detection [10] and face recognition in videos [4], to the best of our knowledge, it has not been explored in the context of face detection in video. In general, the face detection process involves a scanning strategy and a face/non-face classification step. This work introduces a way to improve the face/non-face classification stage by using the spatio-temporal information.

The rest of this paper is organized as follows: Section 2 makes a brief analysis of existing spatio-temporal descriptors. Section 3 describes the Extended set of Volume Local Binary Patterns (EVLBP) descriptor and its application to face detection. The experimental analysis and results are presented in Section 4. Finally, Section 5 concludes the paper and states the future works.

2. Spatio-temporal representations for face description

Different types of features have been proposed to describe and classify faces such as color, texture and biological inspired features [18]. In general, discriminative features are required for being able to represent the unique properties of different faces. In the case of videos, the obtained features should describe both the face and its motion patterns, independently of changes in its appearance and difficult backgrounds.

In the last years, several approaches have been proposed in order to consider not only the spatial but also the temporal information contained in videos [1]. This kind of representation, namely spatio-temporal, encodes both the appearance of faces in the spatial domain as well as the discriminative information related to facial movements; independently of shifts, scales and cluttered backgrounds. Moreover, these spatio-temporal descriptors are usually extracted directly from videos and therefore avoid possible failures of other pre-processing methods such as motion segmentation and tracking. Nevertheless, just a few of them make use of the local appearance of faces for discriminative purpose, which have been shown to be of great importance [6].

There exist many local spatio-temporal descriptors, which are merely extensions of local appearance based features to the video domain. Most of them such as the 3D SIFT [14], the extended SURF [16], the HOG-3D descriptor [9] and the 3D Haar-like volumetric features [8], have been proposed for video applications not related to face analysis. One of the few local spatio-temporal descriptors used in video face recognition, is the generalization of the popular Local Binary Patterns (LBP) descriptor to the spatio-temporal domain [4]. This descriptor is a flexible operator, robust with respect to gray scale changes, rotations and translation.

Despite the fact that local spatio-temporal descriptors have been widely used, to the best of our knowledge, they have not been applied for face detection in video before. In order to analyze whether spatio-temporal descriptors are suitable for discriminating between faces and non-faces in video shots, the Extended set of Volume LBP (EVLBP) descriptor [4] was selected; due to it has been shown to be successful for video face recognition and it could increases the efficiency and the effectiveness in the integration with the face recognition algorithm.

3. Extended set of Volume Local Binary Patterns for detecting faces in videos

The original LBP operator labels the pixels of an image by thresholding its 3x3 neighborhood and considering the result as a binary number. It is a powerful texture descriptor which have been very popular mainly because of its computational simplicity and its demonstrated discriminative properties [13].

Many extensions of LBP have been proposed in the literature for still images analysis [13], and recently, it was extended to the spatio-temporal domain [20]. The Volume LBP (VLBP) [20] considers the video sequence as a rectangular prism (or cube) and defines the local neighboring operations in the 3D space. Hence, it encodes not only the local spatial information but also the temporal information between consecutive frames. Afterwards, the VLBP was improved with the Extended set of VLBP (EVLBP) [4], by using different radius, number of sampling points and sequence intervals. The EVLBP operator can be defined as [4]:

$$EVLBP_{L,(P,Q,S),R} = \sum_{m=0}^{M-1} s(I_{t,m} - I_{t_c,c})2^m, \qquad (1)$$

where t_c is the frame of the center pixel c and t is every frame used on the encoding process. Let L be the time interval between encoded frames, $t = t_c - 2L, t_c - L, t_c, t_c + L, t_c + 2L$. R is the radius for selecting the neighboring pixels: P from frame t_c, Q from frame $t_c \pm L$ and S from frame $t_c \pm 2L$. The total number of compared pixels M is then equal to $P + 2Q + 2S$. $I_{t_c,c}$ is the intensity value of c and $I_{t,m}$ is the intensity value of pixel m in frame t; while $s\{f\} \in \{0, 1\}$ is a thresholding function. An example of the EVLBP encoding process is illustrated in Figure 1 .

Thus, a face sequence is represented by dividing it into several overlapping rectangular prisms and local histograms of EVLBP codes are extracted from each of them. In the case of face detection, we concatenate these local histograms into a single histogram. Then, we apply AdaBoost algorithm [2] for automatically selecting the most discriminative features, discarding the ones related to the redundant information.

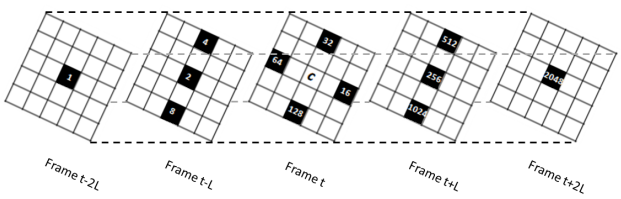


Figure 1. $EVLBP_{2,(4,3,1),2}$ encoding process.

In video face recognition scenarios, it is supposed that faces in each frame are already detected and aligned. Therefore, the information encoded from video sequences certainly contain the target object and any number of frames can be used for the representation. However, in the case of face detection, using a large number of frames can be dangerous because not only the face but also many other information could be encoded. On the other hand, if a very few number of frames are used, some discriminative temporal information can be missed out and more processing time can be required.

For the above reasons, it is necessary to find a trade-off between maximizing the number of frames (N) while minimizing the variability encodes on the descriptor. It is evident that this will be dependent of the video at hand. For example, in a video with a moving face with great variations in shifts, scales and poses, the optimum number of frames to be selected will be much smaller than the number of frames that can be used in an almost static video.

In order to explore the possible values for the number of frames to be selected, an inter-distance matrix between frames were calculated for some face sequences using the L_2 distance. Some frames of used sequences are shown in Figure 2. As can be appreciated they contain faces with variations in expressions, scales and movements in almost static backgrounds, so the distance between different frames is mainly due to the variability on face appearance. In Figure 3 the average distance matrix for 100 sequences is shown. The figure corroborates some expected results such that when picking a small number of frames, the variability is small because closer frames are more similar while far frames can be very different. The average distance matrix suggests that, for example, the smallest variability values are obtained taking around 5 continuous frames (notice that the dark blue area around the diagonal has approximately that width). But this can be considered still a small amount of frames for our purpose. So, we could select instead 10 frames (light blue area around the diagonal) or 15 frames (green area around the diagonal) which are the following noticeable increases in variability.

In real scenario, instead of a common sliding window, we will have a sliding cube of depth N for finding faces

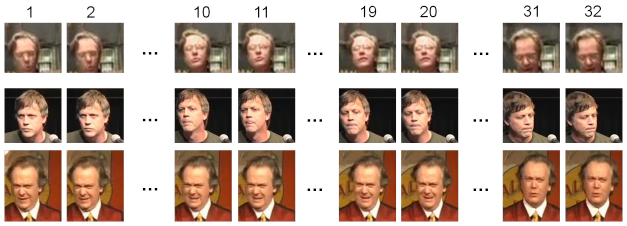


Figure 2. Sample frames from face sequences.

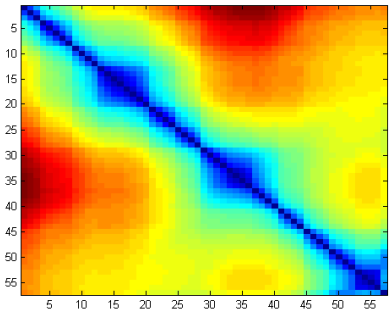


Figure 3. Distance matrix between frames of face video sequences.

over an entire video. For this reason, the greater N, the faster the video will be scanned, taking into account that N should not be arbitrarily large. It should be noticed that once a cube is classified as a face, we will have the same position of the face in each of the N frames.

4. Experimental analysis

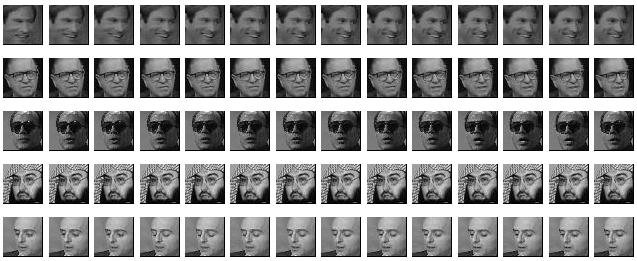
In this section, we aim at corroborating the hypothesis that spatio-temporal representations can improve the face/non-face classification step in the face detection process in video. For this purpose, EVLBP spatio-temporal descriptor is compared to spatial LBP-based representations.

4.1. Experimental protocol

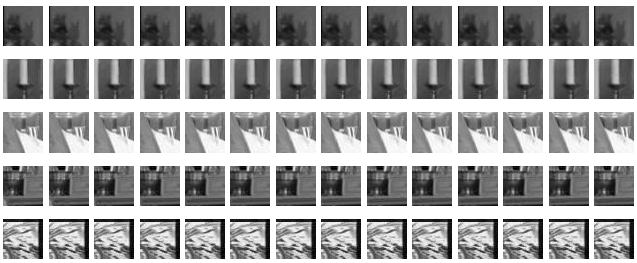
For the experimental analysis we have used the new large and challenging *YouTube Faces* database [17]. Although there exist other video databases containing faces, most of them have been captured under controlled scenarios and with mainly uniform backgrounds. The *YouTube Faces* database contains 3425 videos of 1595 different subjects with significant variations on expression, illumination, pose, resolution, occlusion and background.

The experimental setup used is described as follows: 1091 videos of different subjects were selected, where 500 videos were used for training and 591 for testing. Based on these videos, positive and negative rectangular prisms were built by using N consecutive frames, both for training and testing.

In the case of positives, the annotated position of the face on every frame of the videos is available in the database. The real volume formed by these annotated faces does not necessary form a rectangular prism. Then, the rectangular prism with the greatest intersection with the real volume is used as positive sample. Therefore, the face in each frame is not aligned or cropped, and real displacements of faces in N frames are represented. For negatives examples, the rectangular prisms were obtained from the background area



(a)



(b)

Figure 4. Examples of 14 consecutive frames of some video shots used for training, from (a) face class and (b) non-face class.

of the videos, taking 500 samples for training and 500 for testing. All regions were extracted with 40x40 pixels of size. Some of the obtained positive and negative examples are shown in Figure 4.

In order to extract the EVLBP spatio-temporal descriptor, many different parameter configurations were used, changing the radius, number of sampling points and frames intervals. Based on the empirical study described in Section 3, the selected value for N was 14. Moreover, time intervals between frames was $L = 1, 2$, so just 12 or 10 frames were encoded respectively, which correspond to a suitable value considering the distance matrix shown in Figure 3. As can be seen from Figure 4, there are just little variations between 14 consecutive frames.

4.2. Experimental results

To verify our hypothesis, we evaluate the classification accuracy achieved using EVLBP spatio-temporal descriptor against three different schemes based on spatial LBP-based representations. In all cases AdaBoost classifier was used for selecting and learning the most discriminative features.

The first scheme used is considered as a baseline method, corresponding to a frame by frame classification process, where each frame of the sequence is divided into several local regions from which LBP histograms are extracted and concatenated into an enhanced LBP histogram; and then each of them are independently classified as face/non-face. We refer to this approach as LBP frame-by-frame. In the second scheme, each frame is also represented by a LBP histogram, but the classification results over the face sequence are combined through majority voting. We refer to this approach as LBP + Voting. The third scheme extracts the LBP histogram from each frame and then computes a LBP histogram with the co-occurrences of LBP patterns from all them, which is fed to the classifier. We refer to this approach as LBP + co-occurrence.

Method	FN(%)	FP(%)	ER(%)	CC(%)
EVLBP	**2.54**	**4.21**	**3.30**	**96.70**
LBP frame-by-frame	8.29	12.83	10.37	89.63
LBP + Voting	4.91	10.62	7.52	92.48
LBP + co-occurrence	5.75	9.42	7.43	92.57

Table 1. Comparison between spatial LBP representations and EVLBP spatio-temporal descriptor.

The experimental results are presented in terms of the False Negatives Rate (FN), the False Positives Rate (FP), the total Error Rate (ER), and the total Correct Classification score (CC). In Table 1 the classification performance of each method is listed. As can be appreciated the EVLBP representation exhibits the best accuracy. This means that including the motion information can help to discriminate between faces and non-faces in video sequences. Besides, with this approach, once a subsequence is classified as face, it could be used as an input to a spatio-temporal face recognition algorithm, without losing the temporal order of the frames and at the same time taking advantage of the representation.

5. Conclusions and Future Work

In this work we have evaluated the use of local spatio-temporal representation for discriminating between faces and non-faces in video sequences. Specifically we have selected the EVLBP descriptor, since it has been used for face recognition and we believe that the same feature space can be used for both face detection and classification. When comparing the EVLBP representation with other LBP representations that do not consider the temporal information, it is shown that including the temporal information boost the classification performance.

This paper is just a starting point of this research, a lot of work still needs to be done to have an effective face detector. First we need to integrate the descriptor with a scanning strategy to evaluate and classify every region in the video sequences. For this purpose we are working in the designing of an AdaBoost cascade classifier. Besides, the criteria used for selecting the number of frames encoded in the descriptor could be formulated in a more rigorous way such that the value for this parameter could be obtained as a result of it. Finally, we want to further investigate how can we use the same feature space for detecting and recognizing faces in videos. On the other hand other local spatio-temporal representations can be explored in the context of face detection.

References

[1] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas. Face recognition from video: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(5), 2012.

[2] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

[3] B. Froba and C. Kublbeck. Face tracking by means of continuous detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 5:65, 2004.

[4] A. Hadid and M. Pietikäinen. Combining appearance and motion for face and gender recognition from videos. *Pattern Recognition*, 42(11):2818–2827, nov 2009.

[5] A. Hadid, M. Pietikäinen, and T. Ahonen. A discriminative feature space for detecting and recognizing faces. In *Computer Vision and Pattern Recognition (CVPR)*, pages 797–804, 2004.

[6] B. Heisele, P. Ho, J. Wu, and T. Poggio. Face recognition: component-based versus global approaches. *Computer Vision and Image Understanding*, 91(1-2):6–21, 2003.

[7] O. Javed and M. Shah. Tracking and object classification for automated surveillance. In *European Conference on Computer Vision (ECCV)*, volume 2353, pages 343–357, 2002.

[8] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)*, volume 1, pages 166–173, Washington, DC, USA, 2005. IEEE Computer Society.

[9] A. Kläser, M. Marszaek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference (BMVC)*, 2008.

[10] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In *First International Workshop on Spatial Coherence for Visual Motion Analysis*, 2004.

[11] C. Liu, P. C. Yuen, and G. Qiu. Object motion detection using information theoretic spatio-temporal saliency. *Pattern Recogn.*, 42(11):2897–2906, Nov. 2009.

[12] J. C. Nascimento and J. S. Marques. Performance evaluation for object detection algorithms for video surveillance. In *IEEE Transaction on Multimedia*, 2006.

[13] M. Pietikäinen, A. Hadid, G. Zhao, and T. Ahonen. *Computer Vision Using Local Binary Patterns*. Springer-Verlag London, Ltd, 2011.

[14] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 357–360, New York, NY, USA, 2007. ACM.

[15] H. Wang, Y. Wang, and Y. Cao. Video-based face recognition: A survey. *World Academy of Science, Engineering and Technology*, 36, december 2009.

[16] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 650–663, Berlin, Heidelberg, 2008. Springer-Verlag.

[17] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[18] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song. Recent advances and trends in visual tracking: A review. *Neurocomput.*, 74(18):3823–3831, Nov. 2011.

[19] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical Report MSR-TR-2010-66.

[20] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915 –928, june 2007.