

Combining Motion and Appearance for Gender Classification from Video Sequences

Abdenour Hadid and Matti Pietikäinen

Machine Vision Group, P.O. Box 4500, FI-90014, University of Oulu, Finland

Abstract

We investigate whether combining appearance (face structure) and motion (the way a person is talking and moving his/her facial features) boosts gender classification from face sequences. We propose and compare different schemes based on appearance only, motion only, and combination of appearance and motion. Experiments on various face video datasets of persons uttering phrases or expressing emotions show that combination of motion and appearance is useful for gender analysis of familiar faces, yielding in classification accuracy of 100%. However, for unfamiliar faces, motion seems to not provide additional discriminative information as the best performance (96.3%) is obtained using an appearance based approach with Local Binary Pattern (LBP) features and Support Vector Machines (SVMs).

1 Introduction

Gender classification systems aim to determine whether the person in the given image or video is a man or a woman. Determining such information is useful for many applications such as more affective Human Computer Interaction (HCI), restricting access to certain areas based on gender, collecting demographic information in public places, counting the number of women entering a retail store and so on.

First attempts of using computer vision based techniques to gender classification started in early 1990s. Since then, many approaches have been reported in literature. Among the most notable results to date are those obtained by Moghaddam and Yang [5], and also by Baluja and Rowley [2]. Moghaddam and Yang used raw pixels as inputs to Support Vector Machines (SVMs) and achieved a classification rate of 96.6% on FERET database of images scaled to 12×21 pixels [5]. Note that the considered FERET images were very clean and some persons may have appeared in both training and test sets. Comparable accuracy but at a

higher speed was also reported by Baluja and Rowley who used AdaBoost to combine weak classifiers, constructed using simple pixel comparisons, into single strong classifier [2].

Both approaches cited above are based on still images and assume well aligned faces. However, in many real applications (e.g. HCI and visual surveillance) input data generally consists of video sequences and it is not always obvious to hold the face alignment assumption. One way to enhance the performance of gender classification techniques in such environments is to design multi-modal systems combining different cues such as face, gait, facial dynamics and voice. For instance, Shan *et al.* investigated the fusion of face and gait at feature level and obtained performance increase when combining the two cues [7]. Naturally, in some applications such as HCI, the gait information may not be available. While some researchers have also investigated the combination of face and voice, surprisingly no work has addressed the combination of face appearance and facial dynamics (i.e. the way a person is talking and moving his/her facial features) to gender classification. Perhaps, this issue is understudied mainly because of two reasons: (i) the role of facial dynamics in face analysis is still under debate; and (ii) it is not obvious to define a proper facial representation efficiently combining facial appearance and motion.

Motivated by the psychophysical findings (e.g. [4]) which indicate that facial movements can provide valuable information to gender recognition and also by our recent success of using LBP (Local Binary Patterns) [6] for combining appearance and motion for face and facial expression recognition [8, 3], we adopt in this work the LBP approach to study and address whether combining appearance and motion also boosts the discrimination between men and women. We propose and compare different schemes based on appearance only, motion only, and combination of appearance and motion. We report our experimental results on various face video datasets of persons uttering phrases or expressing emotions.

2 Local Binary Patterns

The LBP operator [6] is defined as a gray-scale invariant texture measure, derived from a general definition of texture in a local neighborhood. It is a powerful means of texture description and among its properties in real-world applications are its discriminative power, computational simplicity and tolerance against monotonic gray-scale changes caused, e.g., by illumination variations. LBP can be efficiently used for representing and analyzing faces in both still images and video sequences [1, 3, 8].

2.1 LBP in Spatial Domain

The original LBP operator forms labels for the image pixels by thresholding the 3×3 neighborhood of each pixel with the center value and considering the result as a binary number. The histogram of these $2^8 = 256$ different labels can then be used as a texture descriptor. The operator has been extended to use neighborhoods of different sizes. The calculation of the LBP labels can be easily done in a single scan through the image. The value of the LBP code of a pixel (x_c, y_c) is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \text{ where } g_c \text{ corresponds to}$$

the gray value of the center pixel (x_c, y_c) , g_p refers to gray values of P equally spaced pixels on a circle of radius R , and s defines a thresholding function as follows:

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0; \\ 0, & \text{otherwise.} \end{cases} \text{ Another extension to the}$$

original operator is the definition of so called *uniform patterns*. This extension was inspired by the fact that some binary patterns occur more commonly in images than others. This yields to the following notation for the LBP operator: $LBP_{P,R}^{u2}$. The subscript represents using the operator in a (P, R) neighborhood. Superscript $u2$ stands for using only uniform patterns and labeling all remaining patterns with a single label. Each bin can be regarded as a micro-texton. Local primitives which are codified by these bins include different types of curved edges, spots, flat areas etc.

2.2 LBP in Spatiotemporal Domain

The original LBP operator was defined to only deal with spatial information. Recently, it has been extended to a spatiotemporal representation for dynamic texture analysis. This has yielded the so called Volume Local Binary Pattern operator (VLBP) [8]. The idea behind VLBP consists of looking at dynamic texture as a set of volumes in the (X, Y, T) space where X and Y denote the spatial coordinates and T denotes the frame

index (time). The neighborhood of each pixel is thus defined in three dimensional space. Then, similarly to LBP in spatial domain, volume textons can be defined and extracted into histograms. Therefore, VLBP combines motion and appearance together to describe dynamic texture. Later, to make the VLBP computationally simple and easy to extend, the co-occurrences of the LBP on three orthogonal planes (LBP-TOP) were also introduced [8]. LBP-TOP consists then of considering three orthogonal planes: XY , XT and YT , and concatenating local binary pattern co-occurrence statistics in these three directions. The circular neighborhoods are generalized to elliptical sampling to fit to the space-time statistics. More recently, a variant of VLBP operator which handles better the temporal information is considered to derive a richer set of volume LBP features denoted EVLBP [3].

3 Gender Recognition from Videos Using LBP

LBP approach has been successfully used for combining motion and appearance in face recognition [3] and also in facial expression recognition from videos [8]. To address the challenging question of whether combining appearance and motion information also helps the discrimination between men and women, we propose and compare three schemes using LBP. The first two approaches use the appearance and motion information solely while the third approach combines both cues through a spatiotemporal representation.

3.1 Appearance Based Approach

An approach to perform appearance based gender recognition from videos is to apply some still image based method to each frame and then combine the results through majority voting which consists of identifying the gender in every frame and then fusing the results. Therefore, we divide each facial image (frame) into several local regions from which LBP histograms are extracted and concatenated into an enhanced feature histogram. Then, we present the results to an SVM classifier for gender classification. The classifier decides on the gender of the face according to the sign of $(\sum_{i=1}^l \alpha_i y_i K(h, h_{t_i}) + b)$, where h_{t_i} is the LBP representation of the training sample t_i , y_i is 1 or -1 depending on whether t_i is a positive or negative sample (male or female), l is the number of samples, b is a scalar (bias), and $K(\cdot, \cdot)$ the second degree polynomial kernel function defined by: $K(z_1, z_2) = (1 + z_1 \cdot z_2)^2$, and α_i are the parameters of the SVM classifier, recovered by solving a quadratic programming problem. Finally, we

combine the recognition results over the face sequence using majority voting. We refer to this approach in our experiments as *LBP+SVM+Voting*. Another approach to perform appearance based gender classification from videos is to consider the XY plane and collect the co-occurrences of LBP patterns from all frames. So, we divide each facial image into several local regions, extract LBP features from corresponding local regions in all frames to obtain one local histogram per local region, and finally normalize and concatenate the local histograms into an enhanced feature histogram which is fed to SVM classifier. We refer to this approach as *XY-LBP+SVM*. The main difference between the two approaches lies in the fact that the first one uses image based technique to each frame and combines the results at decision level while the second approach performs feature fusion from all frames before classification.

3.2 Motion Based Approach

Similarly to the second appearance based approach (i.e. *XY-LBP+SVM*), we derive an LBP based approach which uses only the motion information in the face sequences for gender recognition. The idea consists of considering the YT or XT directions for extracting the LBP features which are then used with SVMs for classification. We refer to such approaches as *XT-LBP+SVM* and *YT-LBP+SVM*.

3.3 Spatiotemporal Based Approach

To combine motion and appearance for gender classification, we consider VLBP approach and collect the co-occurrences of volume LBP patterns in the (X,Y,T) space into a feature histogram [8]. The recognition is done using SVM classifier. We refer to this approaches as *VLBP+SVM*. We also consider the use of extended volume LBP features with AdaBoost learning for constructing a strong classifier from weak classifier based on EVLBP codes [3]. We refer to this approach as *EVLBP+AdaBoost*.

4 Experiments

4.1 Data

For experimental analysis, we considered three different publicly available video face databases (namely CRIM¹, VidTIMIT² and Cohn-Kanade³) containing

¹<http://www.crim.ca>

²<http://users.rsise.anu.edu.au/~conrad/vidtimit/>

³http://vasc.ri.cmu.edu/idb/html/face/facial_expression/index.html

several subjects moving their facial features by uttering phrases, reading broadcast news or expressing emotions. CRIM is a large set of 591 face sequences showing 20 persons (10 female and 10 male) reading broadcast news for a total of about 5 hours. There are between 23 and 47 video sequences for each individual. The VidTIMIT database consists of audio recordings and video sequences of 43 subjects (19 female and 24 male), reciting ten short sentences in three sessions with an average delay of a week between sessions, allowing for appearance and mood changes. Cohn-Kanade database consists of 100 subjects expressing different emotions like anger, disgust, fear, joy, sadness, and surprise. Sixty-five percent of the subjects were female, 15 percent African-American, and three percent Asian or Latino.

We randomly segmented the datasets and extracted over 4 000 video shots of 15 to 300 frames each. From each shot or sequence, we automatically detected the eye positions from the first frame. The determined eye positions are then used to crop the facial area in the whole sequence. Finally we scaled the resulted images into 3 different resolutions: 20×20 , 40×40 and 60×60 pixels.

4.2 Results

For evaluation, we adopted a 5-fold cross validation test scheme by dividing the 4 000 sequences into five groups and using the data from four groups for training and the left group for testing. We repeated this process five times and we report the average classification rates. In most experiments, the features are extracted using the following LBP operators: $LBP_{8,1}^{u2}$, $LBP_{8,2}^{u2}$, $LBP_{4,1}$ and $VLBP_{1,4,1}$ in local regions of sizes from 10×10 to 20×20 pixels. All built SVM classifiers are using second degree polynomial kernel functions.

When dividing the data into training and test sets, we explicitly considered two scenarios. In the first one, a same person may appear in both training and test sets with face sequences completely different in the two sets due to facial expression, lighting, facial pose etc. The goal of this scenario is to analyze the performance of the methods in determining the gender of familiar persons seen under different conditions. In the second scenario, the test set consists only of persons who are not included in the training sets. This is equivalent to train the system on one or more databases and then do evaluation on other (different) databases. The goal of this scenario is to test the generalization ability of the methods to determine the gender of unseen persons.

The gender classification results using all considered methods (motion only, appearance only and combina-

Table 1. Gender classification results on test videos of familiar (columns 1-3) and unfamiliar subjects (columns 4-6). The methods are based on appearance only (1st, 2nd & 3rd rows), motion only (4th & 5th rows), and combination of appearance and motion (6th & 7th rows).

Method	Gender Classification Rate					
	Subjects Seen during Training			Subjects Unseen during Training		
	20×20	40 × 40	60×60	20×20	40 × 40	60×60
Pixels+SVM+Voting	93.1	93.3	91.9	88.5	89.4	88.2
LBP+SVM+Voting	94.0	94.4	95.4	90.1	90.6	91.0
XY-LBP+SVM	96.1	97.2	97.1	95.5	95.7	96.3
YT-LBP+SVM	74.5	81.6	83.2	51.6	49.7	50.4
XT-LBP+SVM	78.5	79.4	80.4	45.9	47.1	44.2
VLBP+SVM	98.2	98.3	98.8	82.7	84.3	84.7
EVLBP+AdaBoost	100	100	100	79.2	81.5	78.6

tion of the two cues) in both scenarios (familiar and unfamiliar) and three different image resolutions are summarized in Table. 1. For comparison, the results obtained using raw pixels with SVM and voting are also shown. We can notice that all methods gave better results with familiar faces than unfamiliar ones. This is not surprising and can be explained by the fact that perhaps the methods did not rely only on gender features for classification but may also exploited information about face identity. The results also indicate that solely motion based methods gave poor results in both scenarios especially with unfamiliar faces. For familiar faces, the combination of appearance and motion yielded in best results, while for unfamiliar faces the best results are obtained using appearance based methods. This indicates that incorporating motion information with appearance was very useful for familiar faces but seems to not provide additional discriminative information with unfamiliar ones. Our results also show that image resolution does not affect very much gender classification performance and this confirms the conclusions of many other researchers. However, handling faces under severe illumination changes and miss alignments needs probably features which are more discriminative than simple raw pixel values.

5 Conclusion

Motivated by the psychophysical findings which indicate that facial movements can provide valuable information to gender recognition and also by our recent success of using LBP for combining appearance and motion for face analysis from videos, we investigated whether combining the two cues (i.e. appearance and motion) also boosts the discrimination be-

tween men and women. We proposed and compared different schemes and our experimental results show that combination of motion and appearance is useful for gender analysis of familiar faces. For unfamiliar faces, motion seems to not provide additional discriminative information. Perhaps, the existing spatiotemporal representations have not yet shown their full potential and need further investigations.

Acknowledgment

The financial support of the Academy of Finland is gratefully acknowledged.

References

- [1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. PAMI*, 28(12):2037–2041, 2006.
- [2] S. Baluja and H. Rowley. Boosting sex identification performance. *IJCV*, 71:111–119, 2007.
- [3] A. Hadid, M. Pietikäinen, and S. Z. Li. Learning personal specific facial dynamics for face recognition from videos. In *AMFG 2007*, pages 1–15, 2007.
- [4] H. Hill and A. Johnston. Categorizing sex and identity from the biological motion of faces. *Current Biology*, 11(11):880–885, 2001.
- [5] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE Trans. PAMI*, 24(5):707–711, 2002.
- [6] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. PAMI*, 24:971–987, 2002.
- [7] C. Shan, S. Gong, and P. McOwan. Learning gender from human gaits and faces. In *AVSS’07*, pages 505–510, 2007.
- [8] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. PAMI*, 29(6):915–928, 2007.