An analysis of facial description in static images and video streams

M. Castrillón-Santana, J. Lorenzo-Navarro, D. Hernández-Sosa, and Y. Rodríguez-Domínguez

IUSIANI
Edif. Ctral. del Parque Científico Tecnológico
Universidad de Las Palmas de Gran Canaria, Spain
mcastrillon@iusiani.ulpgc.es

Abstract. This paper describes an analysis performed for facial description in static images and video streams. The still image context is first analyzed in order to decide the optimal classifier configuration for each problem: gender recognition, race classification, and glasses and moustache presence. These results are later applied to significant samples which are automatically extracted in real-time from video streams achieving promising results in the facial description of 70 individuals by means of gender, race and the presence of glasses and moustache.

1 Introduction

Human beings are sociable by nature and use their sensorial and motor capabilities to communicate with their environment. If Human to Computer Interaction (HCI) were more similar to human to human communication, accessing HCI devices would be easier and this fact would improve their social acceptability, becoming non-intrusive, more natural and comfortable [9].

Among the different channels used for human communication, the face has great importance conveying to humans a wealth of social signals, being therefore considered the center of human communication [6]. They tell us who the person is, or help us to guess features that are interesting for social interaction such as gender, age, expression and more. That ability allows us to react differently with a person based on the information extracted visually from his/her face. For these and other reasons, computer-based facial analysis is becoming widespread, covering applications such as identity recognition, gender recognition, facial expression analysis, etc.

The contribution of this work is the analysis of an appearance based approach for semantic facial description of individuals in static images and during an interactive session. The paper is organized as follows: in Section 2 the approach used for facial description in still images is described and tuned for the problems selected. Section 3 considers the application to video streams, establishing a criteria for pattern selection during interaction. Finally, in Section 4 the main conclusions of the work are outlined, as well as directions for future development.

2 Facial Description

The facial descriptors considered in this work are: gender, race, and the presence or not of moustache and glasses. In the literature different works have tackled the problem of gender recognition. A recent approach based on rincipal Components Analysis (PCA) achieves high performance also for low resolution images [8]. In [7] a Gabor wavelet representation on selected points is used with good results in gender and race classification. There are different references [4, 12] which try to detect the presence of glasses in a face, but we have not found any reference tackling the the problem of the moustache presence.

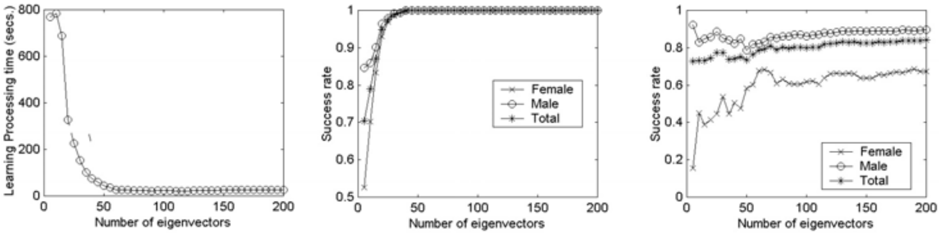

Fig. 1. Gender results: left) Model training time, middle) success rates for training set, and right) success rate for test set.

To tackle the problem, a representation mechanism must first be established to represent faces once the input data, i.e. the images, are available. It is interesting to reduce the data dimensionality to encode the face image without losing information. We selected a well known face representation space in advance: the PCA space due to its economical advantages [5]. The different classifications are performed in that representation space by means of a Support Vector Machines (SVM) classifier [10]. This combination $PCA + SVM$ has been chosen for being well known by the community and the good performance results achieved [2].

The different classifiers performance is analyzed in relation to the number of eigenvalues used for representation, in order to get the best number for reliable classification in each problem. To define the PCA space, we have previously annotated the eye positions of 6000 faces of different people taken from internet. These images have been normalized according to eye positions obtaining 59×65 samples which were used for the gender and race descriptors, and more localized areas to check the presence of glasses and moustache, see Figure 4.A. The PCA space calculation using 4000 of them required 12 hours in a PIV 2.2 Ghz. Different training and test sets have been set up for each problem, see Table 1.

Gender recognition: The results, see Figure 1, show that the training set needs around $40 - 50$ eigenvalues to be perfectly classified, while the test set presents a balanced improvement for both sets up to 70 eigenvalues.

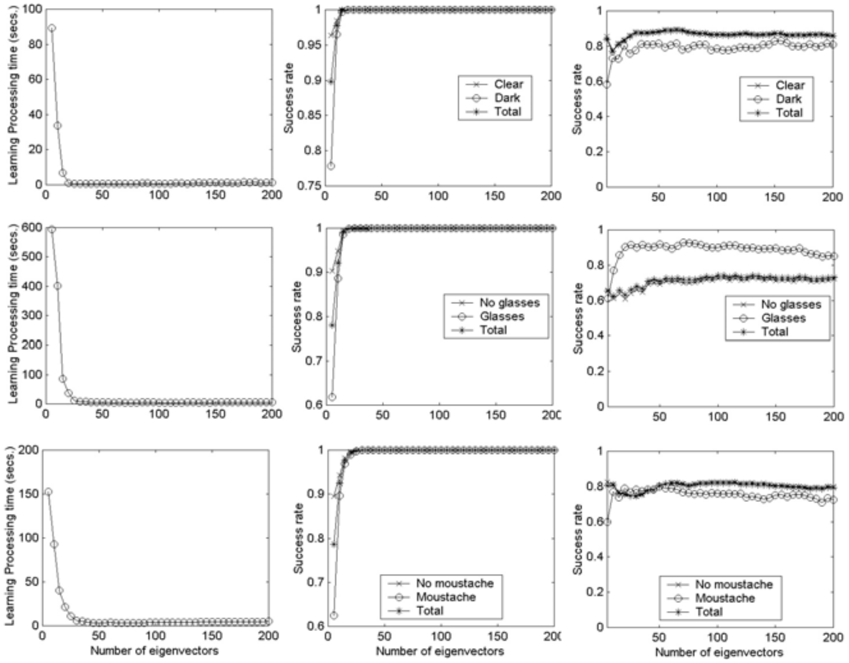

Fig. 2. Each row represents the ratios achieved for the different descriptors: gender, skin, glasses and moustache. In each row left) Model training time, middle) success rates for training set, and right) success rate for test set.

The required training time is also reduced for more than 60 eigenvalues. According to some results on human perception [3], the experiments have been also performed considering only the eyes area, achieving a performance only 5 points lower.

Race classification: Restricted by the face database, we have considered only two race groups, *clear* and *dark*, suffering problems to find more samples for the *dark* class during our gathering stage. For that reason the training set is smaller than the one used for gender recognition. First row in Figure 2 reflects the results achieved. Due to the unbalanced distribution of the test set, the total success rate is close related to the clear skin class rate, however it is observed that around 30 eigenvalues are necessary to classify correctly the training set, while the best results for the test set are achieved in the range $50 - 70$.

Glasses presence: For the glasses presence problem, we have restricted the image to the eyes area, see Figure 4.A. Middle row in Figure 2 reflects the results, the test set is correctly classified with around 30 eigenvalues, while the test set starts to lose some performance (observing both sets) with more than 80 eigenvectors.

	Training set		**Test set**	
Descriptor	Female	Male	Female	Male
Gender	1223	1523	835	2246
Descriptor	Clear	Dark	Clear	Dark
Race	574	316	4811	306
Descriptor	No	Yes	No	Yes
Glasses presence	912	692	4042	356
Moustache presence	710	480	4389	426

Table 1. Training and test sets. We have tried to build balanced training sets, but for some descriptors one class is not so frequent in our database, and therefore the training set is reduced and the test set has much more samples of the most typical class.

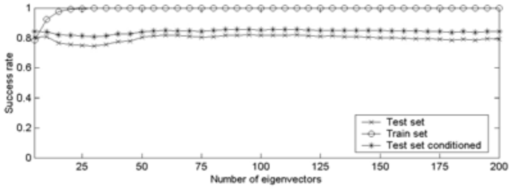

Fig. 3. Moustache conditioned.

Moustache presence: For moustache, bottom graphs in Figure 2 presents the results observing that with more than $50-60$ eigenvalues the test set starts to lose the success rate.

Conditional classification: We have also considered the application of a classifier attending to a previous condition. It is supposed that a female has no moustache, therefore, we apply the moustache presence classifier only if the face was considered male. The results reflected in Figure 3 indicates that this information, with the current success rates achieved, improves the performance for the test set.

According to these results, the optimal number of eigenvalues to use are 70 for gender recognition and glasses presence, 60 for moustache presence, and 50 for race classification. In the next section we analyze their performance processing faces automatically detected in video streams.

3 Video Stream Processing

Our final objective is to be able to provide the system the ability of describing an individual who interacts with, therefore we apply the conclusions extracted in the previous section to video stream analysis.

For that purpose an automatic face detector is required. The one employed combines the general object detection framework by Viola and Jones [11], skin color detection, tracking and temporal coherence providing high performance,

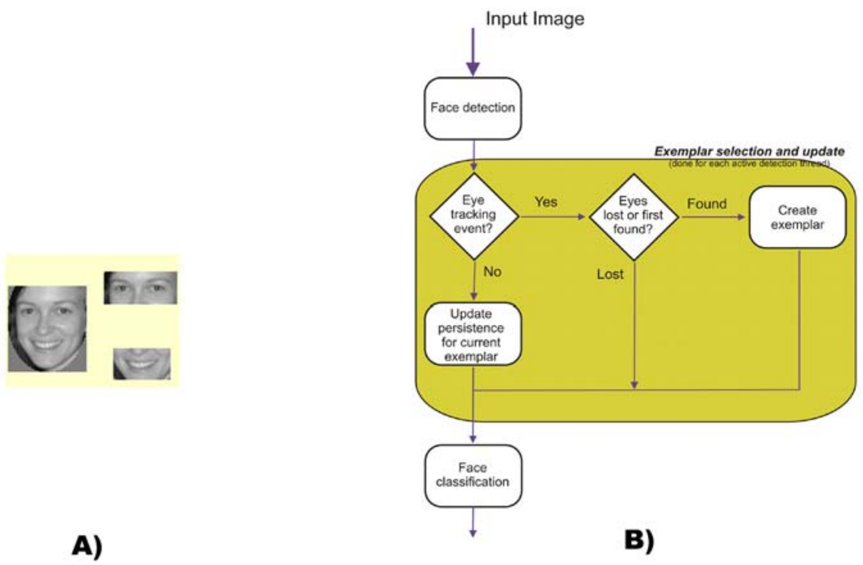
Input Image
Face detection
Exemplar selection and update
(done for each active detection thread)
Eye tracking event?
Yes
Eyes lost or first found?
Found
Create exemplar
No
Lost
Update persistence for current exemplar
Face classification
A)
B)

Fig. 4. A) Images areas used for the different problems. B) Exemplar selection process

see [1] for more details. For each detected face, the system stores not only its position and size, but also its average color and patterns. In summary, each face detected in a frame can be characterized by different features $x_i = \langle pos, size, color, eyes_{pos}, eyes_{pattern}, face_{pattern} \rangle$.

During an interaction session, IS, the face detector gathers a set of detection threads, $IS = \{dt_1, dt_2, ..., dt_n\}$. A detection thread contains a set of continuous detections, i.e. detections which take place in different frames and are related by the system in terms of position, size and pattern matching techniques. Thus, for each detection thread, the face detector system provides a number of facial samples, $dt_p = \{x_1, ..., x_{m_p}\}$.

3.1 Significant patterns selection

As mentioned in the previous section, the face detection system provides a set of detection threads. From each, some selected patterns, the exemplars $e_p = \{e_1, ..., e_{s_p}\}$, are extracted in order to reduce information redundancy.

The criteria used to select significant samples in a detection thread, have been chosen to be easily integrated in the detection process. For that reason, it is based on events reported by the the eye tracker integrated in the face detector, see [1] for more details about that detector. A tracking failure shows an evidence of a substantial change in the face appearance, which forces the tracker to lost the target. Under this circumstance, the system needs to use another cue to detect again first the face and later the eyes, or the detection thread will be considered lost. The first face detected in the next frames by the eye tracker is taken as a new exemplar, see Figure 4.B for a graphical overview of the selection process. For each exemplar, its time life until the next tracking failure is stored. Therefore, an exemplar is described by the data provided by the normalized detected face, x_j, and its persistence, pe_j, i.e. $e_j = \langle x_j, pe_j \rangle$. In Figure 5, the exemplars extracted automatically for two individuals during sessions of more than 15 secs. are presented.

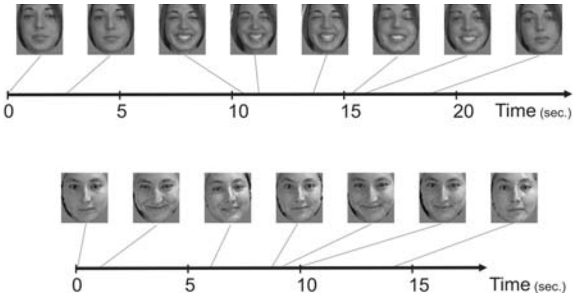

Fig. 5. Stable patterns or exemplars extracted from two different detection threads. Dot lines indicates the moment in which they were extracted during interaction.

Given an interaction session, IS, for any detection thread, dt_p, a facial classifier can compute the likelihood for a class, C_k. This is done by weighting the binary classification for each exemplar according to its relative persistence in relation to the total persistence of the detection thread. This is expressed as:

$$P(C_k|dt_p) = \frac{\sum_{j=1}^{s_p} P(C_k|e_j) * pe_j}{\sum_{n=1}^{s_p} pe_n} \qquad (1)$$

3.2 Experiments with video streams

70 sequences corresponding to different individuals, cameras and environments with a resolution of 320×240 were recorded and processed. The total set contains 27271 images, presenting all of them a face easily detected by a human. The face detector located 98.5% of them with an error rate of 5%.

Table 2 summarizes the results for the different descriptors, computed with (1) for the exemplars automatically extracted from each sequence. The correct classification rates are above 80% for moustache and glasses presence problems. For race classification the results are above 90% for both classes, but it must be noticed that the number of dark individuals is reduced in the test set. For gender recognition the results are worse, over 70% for both sets using the eyes area, and over 65% using the whole face.

This low confidence achieved by the gender recognizer can be used by the system to suggest a classification only if the winner class has a likelihood greater than 0.7, asking for supervision in any other situation. This action will additionally allow the system to distinguish who is not correctly classified, and therefore who should be added to the training set, due to the fact that his/her particular data are still not properly considered in the gender model. That information can be used by the system to tune the classifier based on its experience, in order to learn iteratively a better classifier.

	$P(\neg F) > 0.7$	$P(\neg F) > 0.5$	$P(F) > 0.5$	$P(F) > 0.7$
Male (46), using the face	56.5%	65.2%	34.6%	17.3%
Male (46) using the eyes	65.2%	71.7%	28.2%	15.2%
Female (24) using the face	4.1%	4.1%	95.8%	83.3%
Female (24) using the eyes	8.3%	20.8%	79.1%	66.6%

	$P(C) > 0.7$	$P(C) > 0.5$	$P(\neg C) > 0.5$	$P(\neg C) > 0.7$
Clear skin (67)	89.5%	94%	5.6%	1.4%
Dark skin (3)	0%	0%	100%	33.3%

	$P(\neg G) > 0.7$	$P(\neg G) > 0.5$	$P(G) > 0.5$	$P(G) > 0.7$
No glasses (59)	81.3%	86.4%	13.5%	11.8%
With glasses (11)	9%	18%	81%	36%

	$P(\neg M) > 0.7$	$P(\neg M) > 0.5$	$P(M) > 0.5$	$P(M) > 0.7$
No moustache (64)	92.2%	98.4%	1.5%	0%
With moustache (6)	0%	16.6%	83.3%	66.6%

Table 2. Results achieved for facial description. The left column reflects in brackets the number of individuals (video streams) with a particular feature. The other columns indicate the percentage of those sequences which were labelled with a likelihood of belonging to a class (F for female, $\neg F$ for male, C for clear skin, $\neg C$ for dark skin, G for glasses, $\neg G$ for no glasses, M for moustache and $\neg M$ for no moustache). For example, the value in the second row and column, 56.5%, indicates that this percentage of sequences was assigned to the class $\neg F$, i.e. Male, with a likelihood greater than 0.7.

4 Conclusions

An analysis has been performed for facial description in static images and video streams. A subset of the total number of eigenvectors has been empirically selected in order to get better performance for each problem. An approach for significant samples extraction from video streams has also been described. The results achieved classifying automatically selected faces in video streams of individuals not contained in the training set are decent enough to keep on developing these abilities for a machine.

Further work must focus on gathering more interactive sessions with individuals with features less frequent in our test set, to perform further experiments. Additionally, we are interested in developing some tools for self supervision of the system in order to improve the current classifiers by means of its own experience.

Acknowledgments

Work partially funded by research projects Univ. of Las Palmas de Gran Canaria UNI2003/06, UNI2004/10 and UNI2004/25, Canary Islands Autonomous Government PI2003/160 and PI2003/165 and the Spanish Ministry of Education and Science and FEDER funds (TIN2004-07087).

References

1. M. Castrillón Santana, J. Lorenzo Navarro, O. Déniz Suárez, and A. Falcón Martel. Multiple face detection at different resolutions for perceptual user interfaces. In *2nd Iberian Conference on Pattern Recognition and Image Analysis*, Estoril, Portugal, June 2005.
2. O. Déniz Suárez, M. Castrillón Santana, and F. M. Hernández Tejera. Face recognition using independent component analysis and support vector machines. *Pattern Recognition Letters*, 24(13):2153–2157, September 2003.
3. F. Gosselin and P. G. Schyns. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, pages 2261–2271, 2001.
4. Zhong Jing and Robert Mariani. Glasses detection and extraction by deformable contour. In *International Conference on Pattern Recognition*, 2000.
5. Y. Kirby and L. Sirovich. Application of the karhunen-love procedure for the characterization of human faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1), July 1990.
6. Christine L. Lisetti and Diane J. Schiano. Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive science intersect. *Pragmatics and Cognition (Special Issue on Facial Information Processing: A Multidisciplinary Perspective*, 8(1):185–235, 2000.
7. Michael J. Lyons, Julien Budyneck, and Shigery Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, December 1999.
8. Baback Moghaddam and Ming-Hsuan Yang. Learning gender with support faces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):707–711, 2002.
9. Alex Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 107–119, January 2000.
10. V. Vapnik. *The nature of statistical learning theory.* Springer, New York, 1995.
11. Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, pages 511–518, 2001.
12. Bo Wu, Haizhou Ai, and Ran Liu. Glasses detection by boosting simple wavelet features. In *17th Int. Conf. on Pattern Recognition, Cambridge, UK*, pages 292–295, August 2004.