Face Dynamics for Biometric People Recognition

Marco Paleari [1], Carmelo Velardo [2], Benoit Huet [1], Jean–Luc Dugelay [2]

[1] *Video–Analysis group,* [2] *Image Security and Biometric group*
Multimedia Department
EURECOM
Sophia Antipolis, France
[last_name]@eurecom.fr

***Abstract*—Biometric systems have gained the attention of both the research community and the industry becoming an important topic in real application scenarios. Face recognition is, with fingerprint, among the most used techniques since it is natural for humans to recognize people from facial appearance, since the technology is mature, and because, unlike fingerprint, it is completely unintrusive. Existing systems only focus on the appearance of the subjects considering facial expressions as an obstacle to their aim. On the other hand such systems presents several limitations when dealing with variable illumination conditions, head pose, day–to–day variations (e.g. beard, glasses, or make–up), etc. Furthermore, most of the current techniques do not exploit dynamics to detect the liveness of the tested subjects. In this paper we present a study on person recognition from the dynamics of the facial feature points. The aim of this work is to demonstrate that dynamics of facial expressions could be seen as a biometric characteristic. Therefore, only dynamic characteristics are considered and the adopted features are purged of all appearance information. The results clearly show that relevant biometric information can be extracted from facial expressions and other dynamics of the face.**

I. Introduction

Over the last decades the market of biometric person recognition has gained the attention of both the research community and private investors.More and more private investors and public administrations are asking for systems capable of automatically and unintrusively recognize people to guarantee security of people, objects, and sensible data.

Although biometric solutions are becoming more and more appealing and mature, unconstrained biometric identification remain a largely unsolved problem. State of the art recognition systems are still far away from the capability of the human perception system.

Nowadays, the most used biometric feature in commercial products is fingerprint. Nevertheless, using such a technique carries two substantial drawbacks: need for contact and need for subject cooperation.

Although several possible biometric information can be extracted unintrusively (e.g. voice, gait and stride, and others soft biometrics traits), automatic computer based face recognition is by far the most studied technique.

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00

Most face recognition techniques can be classified into two categories according to the fact that they input still images or video shots. The system belonging to the first category attempt to recognize a subject exploiting only the physiological appearance of the subject; the ones belonging to the second class couple the information about the physiological appearance with information about the dynamic changes of this characteristic over time.

Traditionally, systems dealing with face recognition have to cope with four main challenges: 1) illumination changes, 2) head pose, 3) facial expressions, and 4) variations in facial appearance (e.g. make–up, glasses, or beard).

Eigen–faces [1], is the most used technique in face recognition from still images. Eigenfaces creates a lower dimensional space using the faces of the train base and a standard dimensionality reduction technique, such it is the principal component analysis (PCA). Eigenfaces recognize a test subject by projecting the test image in the same space and finding the subject linked to nearest train image.

Several techniques have been created to deal with head pose and facial expressions. A first one is known as of *Active Appearance Models* (AAM) [2]. In AAMs a robust representation of the subject is obtained by using a small set of parameters based on characteristics extracted from the position of a set of landmarks on the input image.

A second important contribution is represented by the *Elastic Graph Matching* (EGM) and the *Elastic Bunch Graph Matching* (EBGM) techniques. [3], [4] introduce the use of a graph template for the extraction of features. In these techniques a grid template is stretched and deformed to imitate in the best possible way the face in the image. Negative weights and constraints are applied to the deformations of the template in order to set physical boundaries to the deformations of the facial appearance. In the case of EBGM [4] the regular grid is replaced by a 3D graph representation of the human face accordingly to a set of fiducial points.

A first technique involving temporal information make use of *Hidden Markov Models* (HMM) [5]. HMM by their nature represent the temporal characteristics of signal by modeling the different states which better represent the signal in time, and the probabilities to pass from one state to another. In the face recognition state of the art these signals are represented

by either the raw pixel values, Eigen coefficients, or discrete cosine transform coefficients [5].

Perronin et al. [6] use an approach based on HMM and EGM to model differences between couple of images of the same subject. In this work the features represent both the facial appearance and the grid transformations.

Finally, Chen et al. [7] exploited the technique known as *Optical Flow* (OF). In OFs a regular grid of image pixel blocks is tracked all along the video. The result is a grid of movement vectors representing the movement of the face inside the video.

In recent years NIST [8] has promoted an evaluation campaign for facial recognition systems. Different facial expression were depicted as well as different poses and illuminations. Only two different facial expressions (neutral and smiling) were involved in the challenge but we know that the complexity of the human facial expressions is much higher [9].

Although face images and videos are used by several works for biometric people recognition, only few [7] tried to exploit the facial expressions and the facial dynamics themselves as a source of biometric information. Indeed, most of the works in the state of the art still considers facial expressions as noise for the appearance recognition and, therefore, as a characteristic to avoid or suppress.

In this work we aim at demonstrating that the dynamics of emotional facial expressions and speech production (i.e. movements of the lips region), can be seen as a biometric source of information. In other words, for this preliminary study, we make the hypothesis that, exploiting dynamics of facial expressions, we can perform better than random identification.

We point out that emotional facial expressions are not dependent to age [10] and therefore represent a stable source of information over the years. Furthermore, recognition systems based on facial dynamics have the advantage of being robust to illumination changes and variations of the facial appearance (e.g. beard, glasses, make–up).

This work takes direct inspiration from our system [11] for emotional facial expression recognition and make use of the eNTERFACE'05 multimodal emotional database [12].

II. System

In this section we briefly overview our solution for the analysis of facial expressions (for further details please refers to [11]. This approach consists in fusing the information coming from the tracking of some feature points (FP) placed in semantically meaningful locations of the subject's face.

We took inspirations from the results of the study by Ekman and Friesen [10] which assessed that some emotional facial expressions (i.e. anger, disgust, fear, happiness, sadness, and surprise) are independent from sex, ethnicity, age, and culture. We have then demonstrated [11] that emotions can be recognized following the dynamic evolution of a facial expressions. With this paper we assume that the dynamics of facial expressions can be modulated by mainly two components: a first contribute is given by the emotion itself, a second one is mainly related to the subject. If this hypothesis is confirmed, then 1) we could extract biometric information from the dynamics of facial expressions; 2) by testing on the same emotion used for training a higher accuracy should be reached; 3) by testing on a different emotion than the one used for training it should still be possible to recognize the subject.

The dynamic analysis is performed for these six expressions in real-time exploiting the tracking of the displacement of the FP. Those are automatically detected and tracked on the subject's face we want to identify.

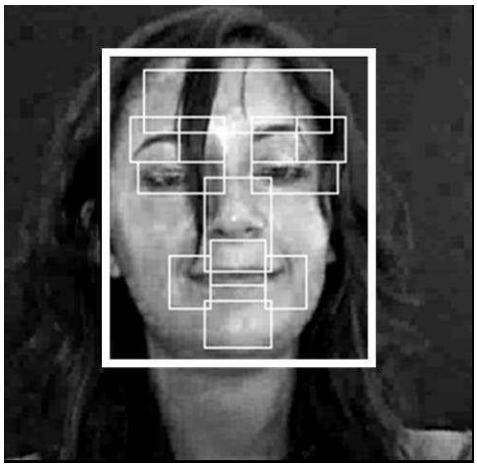


Fig. 1. Anthropometric 2D model

We made use of the eNTERFACE'05 database [12] for our experiments (see figure 1 for an example). The database is a collection of more than 1300 videos regarding 44 non–native English–speaking subjects showing emotions while uttering English sentences. Each sentence is related to one of the six universal emotions from Ekman and Friesen's studies [9]; therefore each video shot represent alternatively one emotion among anger, fear, disgust, happiness, sadness, or surprise; eNTERFACE'05 does not provide neutral expression samples.

Video shot duration is not constant and ranges between 1.2 to 6.7 seconds (2.8 ± 0.8 sec). The eNTERFACE'05 database is publicly available on the Internet but presents some drawbacks in terms of quality because of several factors, some implicit to the representation of the videos and some to the set up of the database. Compression and interlacing applied to the videos, non–professional performances of the actors, and not English–proficiency of the actors (possibly affecting the quality of the vocal expression) affect the quality of the database [11]. The reference paper itself [12] admits that some videos are not fully representative of the related emotions. Dealing with the imperfections of such a database raised some difficulties but we argue that some of them are similar to the ones that can be found in real application scenarios. Therefore we can look at these imperfections as opportunities for devising more reliable methods.

A. Expression analysis

In this section we overview the steps needed for the extraction of facial feature points. We emphasize that facial appear-

ance is not considered in the construction of the features, in fact this framework was conceived for emotion recognition. In our original system the facial appearance is actually a source of noise to the main information: the emotion.

The task of our emotion recognition system is of identifying the six prototypical emotions [10] by fusing the information coming from both visual and audio features. In this study we discard information coming from audio for concentrating on facial expressions.

The system starts identifying the position of the subject face in the shot exploiting the Viola–Jones face detector [13]. For this purpose three different detectors are used. Face, mouth, and eyes of a subject are found in a video shot and exploited to estimate the pose and the face orientation angle.

Once the template of the face is found, we proceed by superimposing an anthropometric model of the human face (see figure 1). For the construction of such a model the distance between the eyes is taken as basic distance.

Using this quasi–rigid 2D model we are able to identify 12 regions of interest (ROIs) on the face target.

These regions of interest are representative of the following face parts (see figure 2(a)):

1) right mouth corner
2) left mouth corner
3) nose
4) right eye
5) left eye
6) forehead
7) mouth bottom / chin
8) external right eyebrow
9) internal right eyebrow
10) internal left eyebrow
11) external left eyebrow
12) upper lip / mouth top

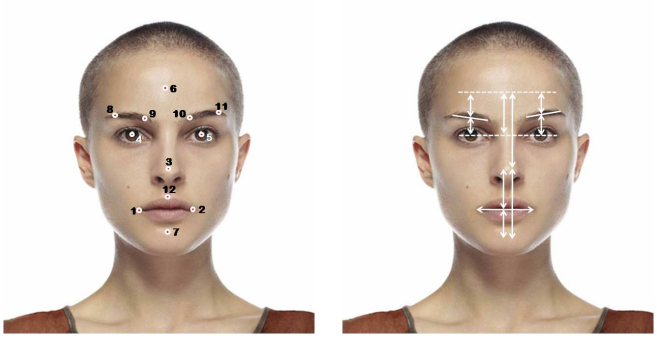
Those regions identify parts of human face that are involved in the emotion creation and expression as verified by [10]. For each one of these ROIs a cloud of Lucas–Kanade [14] points is searched. The center of mass of these points is tracked in time.

The result of such a computation is a pair of $x(i)$ and $y(i)$ coordinates where i represents the corresponding region in the image. Twenty four values (points coordinates) per frame are found (see figure 2(a)). When considered in time, these values represent the average x and y movements of the corresponding ROI.

These coordinates are used as features representing the movement of the face region belonging to the 12 ROIs we identified with the anthropometric template.

Then, we proceed to the normalization of each coordinate with respect to the nose position. This is done to get rid of the dynamics of the head movement. We keep this information aside, stored in the two variables relative to the nose.

We derive here a more meaningful feature set from the 24 signals of the spatial coordinates which better represents the facial dynamics. This new feature set will be completely bound to facial movements.



(a) Feature Points (b) Distances

Fig. 2. Video Features

These new features are computed as distances or alignments among different points (see figure 2(b)).

In this step we reduce the dimensionality of the feature set (from 24 x and y values to 14 values) while keeping the important information about interactions and movements of emotion relevant ROIs. We remove what we consider as redundant information that does not contribute to describe dynamics evolution of face.

The work of deriving this second feature set is directly inspired to the one done with MPEG-4 Face Definition Parameters (*FDPs*) and Face Animation Parameters (*FAPs*). As already demonstrated in a previous work [11] reducing the feature set from coordinates to distances improve both the recognition and the time performances of the algorithms.

The reduced list of features is:

1) head x displacement
2) head y displacement
3) normalization factor proportional to head z displacement
4) mouth corner distance
5) chin distance to mouth
6) nose distance to mouth
7) nose distance to chin
8) left eye to eyebrow distance
9) right eye to eyebrow distance
10) left eyebrow alignment
11) right eyebrow alignment
12) left eyebrow to forehead distance
13) right eyebrow to forehead distance
14) forehead to eye line distance

We extract and isolate the global motion information in three separated variables. This information is representative, as already demonstrated by Matta [15], and can be efficiently exploited for biometric recognition.

For each subject we compute averages for the 14 distances. Then, we proceed to remove those values from each distance signal. The results are then normalized in the range $[-1, 1]$. Thanks to this normalization, we can get rid of the component of information which is linked to the appearance of the subject, thus building up a feature set which only represents the dynamics of the facial expressions.

III. Experimental results and analysis

In this section we are going to present results of our experiments. The tests were conducted using an approach that make use of *GMM*s. We have also tested approaches based on *HMM*s but the limited gap in terms of performances does not justify the increased complexity of this second technique. We argue that such a result is motivated by the not satisfying size of the dataset. The *GMM* used are characterized by one mixture of Gaussians (*MOG*). We also tested an increased number of *MOG* that resulted in comparable performances. Once again this could be due to the relatively small size of the database.

For each pair of *subject* and *expression* the database contains five repetitions. Each repetition represents the same expression but the pronounced words change.

Our results were found by performing leave–one–out tests among the five different sentences.

Three approaches were explored:

1) for each subject we have trained six different *GMM*s (one per emotional facial expression). The tests were carried out using the same expressions;
2) without changing the *GMM*s trained in the first step, we tested using data coming from different facial expressions (e.g. training on anger and testing on fear, disgust, etc.);
3) for each subject a single *GMM* was computed mixing all the data available. Similarly, tests were carried using all the available data.

The results are presented in figures 3 and 4.

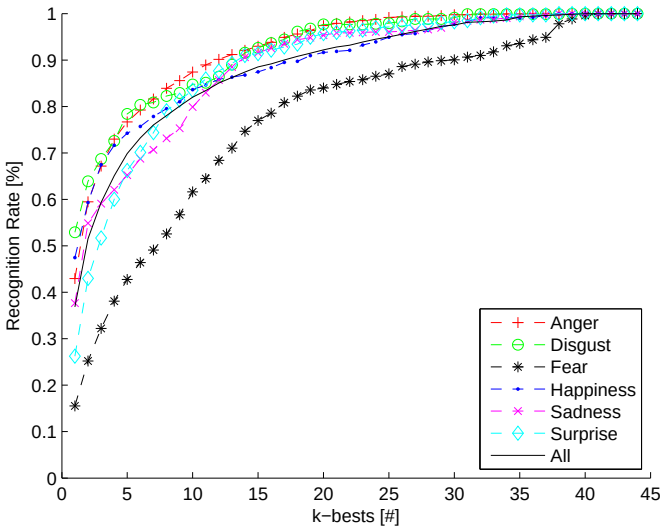


Fig. 3. GMM trained and tested on the same emotional data

Clearly one see from both figures that facial dynamics carry biometric information about the identity of the subjects.

In figure 3 we show the results of the system of *GMM*s trained and tested on a specific emotional dynamics (first approach) and we compare it with the result obtained from the analysis conducted without the emotional state information (third approach). Considering the number of subjects in eNTERFACE'05 database, the average 1–best recognition accuracy is 16 times better than random; in the worst case we perform 7 times better than random.

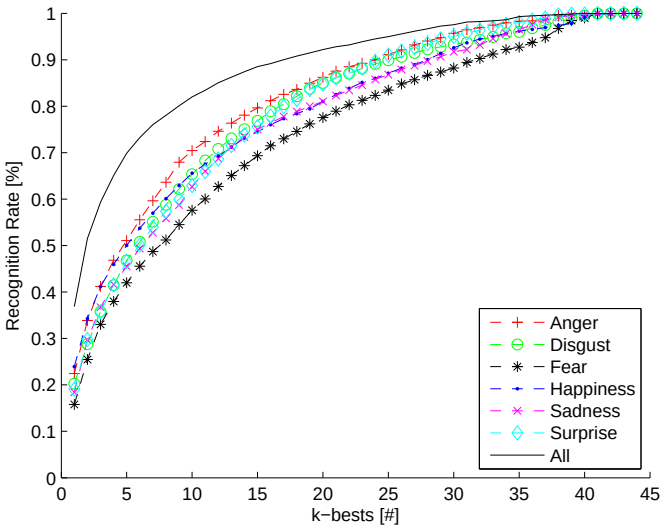


Fig. 4. GMM trained and tested on different emotional data

We can observe that emotion specific *GMM*s perform, in average, sightly better than the output of the mixed approach. Although the distance among curves is small, we point out that the training data size for emotion specific approach is six time smaller than the amount used for the latter. Given the size of our dataset, we believe that such a disparity could affect the results. In other words increasing the dataset size for the emotion specific *GMM*s to the one used for training the mixed approach may further improve the first results.

Another conclusion is that some emotional expressions can better discriminate the distances among subjects. In particular, when subjects depict fear their recognition became harder to accomplish. The reasons for that may be two: from one hand 1) it might be that all the subject have a similar emotional dynamics, on the other hand 2) it could be that the way a subject represent its fear, changes a lot from one video to the other. In the first case we will observe that the intra–subject standard deviations (STD) of the *GMM*s mean values is low; in the second possibility the infra–subject STD of the *GMM*s σ will have low values. From our analysis on the *GMM*s mean and σ values, we can conclude that the latter possibility is verified.

In figure 4 we show the results of the system of *GMM*s trained over an emotional facial expression and tested on the others (second approach). To help the comparison of the results we superimpose here the curve obtained from the analysis conducted without the emotional state information (third approach).

Notwithstanding the fact that performances deteriorated (1–best average classification is halved), we observe that part of

the biometric information is kept (the worst case still performs 7 times better than random).

This verifies the initial hypothesis that two different components are carried by the facial dynamics: one first component being represented through emotional facial expressions and a second one being specific of the subject facial dynamics in term of vocal production, twitches, muscles interactions, etc.

IV. Conclusions

We have shown a system for biometric people recognition based on facial dynamics. We have extracted these characteristics via a robust, automatic, and real–time point tracker. We have demonstrated that emotional facial expressions, up to now considered as noise by the state of the art, carry enough biometric information to distinguish among different people.

With our analysis we have demonstrated that:

1) facial dynamics carry biometric information
2) two different contributions participate to the recognition:
    a) emotional facial expressions
    b) subject dependent dynamics

We point out that algorithms exploiting dynamics are less prone to problems due to illumination and day to day facial variations (make–up, glasses, beard, ...). Furthermore, the dynamics of emotional facial expressions are known to be independent to age, sex, ethnicity, and culture [10]. Therefore, using such characteristics help to build robust and reliable systems.

Additionally, systems exploiting dynamics are less sensitive to spoofing and represent a good tool for capturing the liveness of the tested subject.

We strongly believe that with this work we open a new research path in the study of emotional facial dynamics. Nevertheless, as previously pointed out, a further analysis should be conducted when more data will be available.

References

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[2] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *Computer Vision*, vol. 2, pp. 484–498, 1998.

[3] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. Malsburg, R. Wurtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on Computers*, vol. 42, pp. 300–311, Mar. 1993.

[4] L. Wiskott, J. Fellous, N. Kruger, and C. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 775–779, Jul. 1997.

[5] S. Huang and M. Trivedi, "Streaming face recognition using multicamera video arrays," in *Proceedings of Pattern Recognition*, 2002, pp. 213–216.

[6] F. Perronnin, J.-L. Dugelay, and K. W. Rose, "A probabilistic model of face mapping with local transformations and its application to person recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Volume 27, Issue 7, July 2005*, 2005.

[7] L. Chen, H. Liao, and J. Lin, "Person identification using facial motion," in *Proceedings on Image Processing*, Oct. 2001, pp. 677–680.

[8] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "Frvt 2006 and ice 2006 large-scale experimental results," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. 1, 2007.

[9] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings.* New York: Pergamon Press., 1972.

[10] P. Ekman and W. V. Friesen, "A new pan cultural facial expression of emotion," *Motivation and Emotion*, vol. 10(2), pp. 159–168, 1986.

[11] M. Paleari, R. Benmokhtar, and B. Huet, "Evidence theory based multimodal emotion recognition," in *MMM '09* 15^th *International Conference on MultiMedia Modeling*. Sophia Antipolis, France: ACM, January 2009.

[12] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006.

[13] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. 511–518, 2001.

[14] C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon University, Tech. Rep. CMU-CS-91-132, April 1991.

[15] F. Matta and J.-L. Dugelay, "A behavioural approach to person recognition," in *ICME 2006, IEEE International Conference on Multimedia & Expo, July 9-12, 2006, Toronto, Canada*, Jul 2006.