Face recognition from a *tabula rasa* perspective

M. Castrillón-Santana, O. Déniz-Suárez, J. Lorenzo-Navarro and M. Hernández-Tejera
IUSIANI - Edif. Ctral. del Parque Científico Tecnológico
Universidad de Las Palmas de Gran Canaria, Spain
mcastrillon@iusiani.ulpgc.es

Abstract

In this paper a system for face recognition from a tabula rasa (i.e. blank slate) perspective is described. A priori, the system has the only ability to detect automatically faces and represent them in a space of reduced dimension. Later, the system is exposed to over 400 *different identities, observing its recognition performance evolution. The preliminary results achieved indicate on the one side that the system is able to reject most of unknown individuals after an initialization stage. On the other side the ability to recognize known individuals (or revisitors) is still far from being reliable. However, the observation of the recognition evolution results for individuals frequently met suggests that the more meetings are held, the lower recognition error is achieved.*

1. Introduction

Hundred of approaches [9, 22, 30] have been described for the face recognition problem. Most of them have been designed for the still image context and rarely for video streams [19], despite recent developments in face detection techniques. Generally, it is thought that the best way to test face recognition algorithms is by showing their performance for large numbers of individuals.

A well known corpus used to evaluate recognition techniques is the FERET database [20] and more recently the Face Recognition Vendor Test or the Face Recognition Grand Challenge [19]. This database offers a large enough problem in terms of individuals and samples, but the video context is not considered. Verification approaches make use of the BANCA protocol [3] which tackles the video problem for 208 individuals.

However, it is not proven that the results achieved with those databases can be extended to the whole face domain. Unfortunately, that is the only comparison technique used thus far, even when these systems are still not comparable to human performance in most cases [1, 6].

Our aim is to design a system valid for the video-stream context, i.e. not necessarily high resolution images. A challenge stated at the end of the 90s was to outperform human levels of performance in low-quality images where facial identities seem to be available [7]. This ability is crucial for more natural and comfortable Human Computer Interaction (HCI) [18]. Any Vision Based Interface [24] must include face analysis in order to perceive the user in a HCI context. Therefore, it is assumed for these interfaces that a camera is continuously acquiring images, which can of course register individuals close to the system. In that context, where non invasive techniques are required for facial description, typical approaches are inappropriate [15, 23]. Moreover, the large number of faces collected by the face detector must be processed considering temporal coherence, i.e. the representation and/or classification of individuals should be evaluated in time rather than using an one-shot methodology.

The ability to recognize familiar faces at low resolution is impressive in humans [7, 6]. However, most face recognition approaches tackle the problem using a single image per individual to recognize a large pool of identities [19]. These systems are trying to recognize faces which are not familiar enough becoming less reliable with uncontrolled imagery. Humans are not so reliable for this task, in experiments where the photo ID was not enough to avoid fraud in high levels of performance [13, 7, 21, 6].

We focus the problem of online learning of a face recognition classifier during the system live performance. This is not the case of most classifiers employed in the literature, which were computed off-line and later analyzed with independent, and relatively large, test sets. Can the learning process be done online with current technology? Can the system select from its interactive sessions the info needed to first create and later update the different classifiers according to its experience? In this paper, we describe an approach trying to face this purpose.

1.1. Previous Work

Video stream analysis presents a major difference in relation to still image processing: Individuals present variations along the image stream. In this context it is hard to tackle the face recognition problem based on a single image per individual. The crucial point here is that the ratio of intraclass (similarity between images of faces of a the same individual) to interclass (similarity between images of faces of different individuals) variation is still very high in face recognition, even for a low number of individuals.

Indeed, an object model seems to require a collection of images similarly to the way the human system does [28]. The source to set up such a collection is the interactive sessions that an automatic system has with the particular object. Focusing on face analysis, it is not reliable to use all the images present in a video stream because there is redundancy contained in them, which would produce massive computational and storage costs [2, 29].

The extraction of significant patterns, or exemplars, is tackled in [15], where they are selected from a single gallery video of each individual. However, no further tuning is performed later during classification of new videos. That approach had the novelty of integrating temporal information in the classifier output but did not alter the classifier by means of system experience.

The automatic selection of keyframes, in authors language, used in in [29], is based on tracking failures. That circumstance indicated that a new keyframe should be added to the representational database. Later each new keyframe found during interaction would be compared with those already contained in an individual description and added if needed. This action required robust recognition.

In [2] the authors implemented in a humanoid robot the ability to learn to recognize the people it interacts with. As a novelty, the system was launched with an empty database, exactly the problem that we tackle, and developed a completely unsupervised face recognition system. The system used the standard eigenface method [25], distinguishing two stages: 1) an initial stage where the system must be able to cluster its visual stimuli, and 2) online training, which based the recognition of unknown individuals on a simple distance measure with already stored ones. The detection of an unknown individual allowed the system to create a new identity cluster. In a reduced set of 9 individuals, the system was unable to learn 5 of them using the unsupervised mechanism. The authors affirm that this fact is due to the known performances degradations of the eigenface approach for facial expressions, facial alignment and scale.

Modified Probabilistic Neural Networks are used in [11], being able to identify not only known, but also unknown subjects. Once the system detected an unknown subject, a fixed number of images in the buffer were selected to create new links in the Neural Network. These images were selected according to the difference with the average face computed during the interaction. Once a new model is learned, it will not be updated later.

Exemplars are taken from single images and aligned by hand in [6]. They are later use to compute the average image per identity which is later used to learn. However, the process is not completely automatic, and errors are not used to retrain the system.

2. Extracting Exemplars from Video Streams

2.1. Automatic Face Detection

The real-time face detector, see [8] for more details, combines different techniques providing robust performance in different conditions and environments. An initial detection is obtained by means of window shift detectors [27, 16]. The skin color blob is estimated, and its location used to detect eye positions for frontal faces. Later, temporal coherence is used and each detected face is parameterized in terms of not only its position and size, but also its average color $x_i = \langle pos, size, color, eyes_{pos}, eyes_{pattern}, face_{pattern} \rangle$. These features direct different cues in the next frames which are applied opportunistically in an order based on computational cost and reliability:

- Eye tracking: A fast tracking algorithm [12] is applied in an area that surrounds previously detected eyes, if available.
- Face detector: The Viola-Jones face detector [27] is applied in an area that covers the previous detection.
- Local context face detector: If previous techniques fail, it is applied in an area that includes the previous detection [16].
- Skin color: Skin color is searched in the window that contains the previous detection, and the new sizes and positions are coherently checked.
- Face tracking: If everything else fails, the prerecorded face pattern is searched in an area that covers previous detection [12].

If the eyes are detected, the face is normalized to a 59×65 size. In absence of detections, the process will be based on the standard window shift detectors [27, 16].

2.2. Exemplars Selection

During a meeting with an individual the system relates consecutive detections in terms of position, size and pat-

tern matching techniques, conforming what we call a detection thread, dt. Thus, for each detection thread, the face detector system provides a number of facial samples, $dt_p = \{x_1, ..., x_{m_p}\}$. The presence of gaps or multiple individuals will produce multiple detection threads for a meeting. In order to reduce the huge amount of data extracted during an interactive session, some selected patterns, the exemplars $e_p = \{e_1, ..., e_{s_p}\}$, are extracted for each detection thread, dt_p.

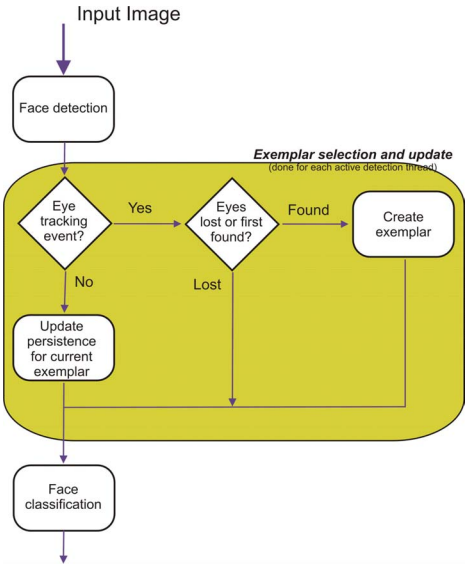

Figure 1. Detailed view of exemplar selection.

The criteria for selecting exemplars, have been chosen in order to be easily integrated in the detection process. Similarly to [29] it is based on tracker events of the face detection system. An eye tracking failure evidences a substantial change in face appearance, forced by a lost target. The system needs to use another cue to detect again first the face and later the eyes. The first face detected and tracked will be taken as a new exemplar. A graphical description is presented in Figure 1. For each exemplar, its time life or persistence until the next tracking failure is stored. Therefore, an exemplar is described by the data provided by the normalized detected face, x_j, its persistence, pe_j, and its timestamp, t_j, i.e. $e_j = \langle x_j, pe_j, t_j \rangle$.

Given an interactive session, IS, for any old enough detection thread (older than 20 frames), dt_p, any facial classifier being considered by the system can compute the *a posteriori* probability for a class, C_k. This is done by weighting the binary classification for each exemplar according to the relative persistence in relation to the total persistence of the detection thread. This is expressed as:

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

Therefore, likely class is suggested for each detection thread computed for the exemplars extracted during the whole interactive session, or for those which have been selected within a recent Window Of Attention (WOA). In that case, in Equation 1 only those exemplars inside the WOA will be considered.

3. Recognition vs. Verification

There are two different problems that share similar techniques in the face identification literature. The first one is associated to recognition from a database without a priori knowledge of the person's identity. The second problem is related to verification or authentication of an identity given by a subject, see Figure 2.

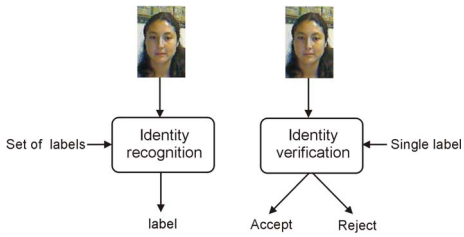

Figure 2. Recognition (left) and Verification Schemas (right).

The first problem is tackled by means of a single n-class classifier that assigns a label to any new image analyzed by the system. The classifier is learnt from a training set which contains samples of those n individuals. If a face image of an individual not contained in the training set is processed, the system is not able to observe that circumstance, it will provide in any case one of those n labels. For the second problem, the literature offers the verification approach to confirm a given identity. Given n identities, the verification system needs n binary classifiers, i.e. a rejection class for each individual, in order to accept or reject the label provided by the user for the face image. These systems are mainly focused on confirming the label provided, but do not guess if the identity is not contained in the database.

To overcome the drawbacks of both systems, and to model the rejection class with available data, we decided to apply both approaches in a cascade manner. The identity classifier has the drawback of not being able to verify if the user is contained in the training set. That can be achieved by a verification stage if a label is provided. Thus, the label

provided by the identity classifier is used for the verification stage, see Figure 3.

This approach forces the system to have a classifier for n classes for the first stage, and n binary classifiers for verification in the second stage.

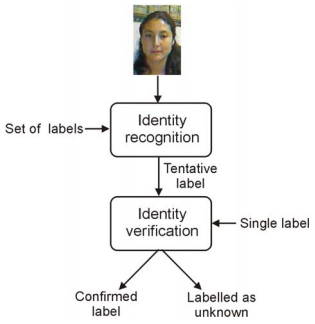

Figure 3. Identity recognition plus verification.

3.1. Representation Space and Classification

Face images have a high dimensionality, feature that makes the classification problem hardly tractable. In order to avoid this problem, Principal Components Analysis (PCA) decomposition [14] is applied to the training data provided. This action allows us to represent the appearance of the different individuals contained in the training set [25].

Using this representation space, different classifiers can be used to select a label for each face processed. The original implementation [25] makes use of Nearest Neighbor Classifier (NCC) for that purpose. However, different authors argued that this approach provides low reliability if lighting conditions are not restricted [5].

Recent developments use local representations such as Independent Components Analysis (ICA) [4] to get a better representation space. However, the work described in [10] proved that the selection of a powerful classification criteria was more critical than the representation space (PCA or ICA). According to these results, recognition experiments have been carried out using Support Vector Machines (SVMs) [26] as classification criteria.

The basic idea for updating the classifiers after a meeting is to make use of the incorrectly classified patterns. At the beginning an expert is needed, similarly to the way humans. Once the system is reliable, there are other mechanisms which will allow the system to learn during its *life* once it is not supervised. For identity recognition it is crucial to detect unknown individuals [2, 11], i.e. individuals which are not already contained in the classifier, in order to create a new identity class.

Wrongly classified exemplars are used to retrain the classifier. For example, if the system were corrected by the expert, and the correct class were C_c, all the incorrectly labelled exemplars, i. e. $P(C_c|e_j) = 0$, will be added to the classifier iteratively whenever the recomputed classifier keeps classifying them wrongly. If the supervisor confirmed the class suggested, C_k, similarly incorrectly assigned exemplars, $P(C_k|e_j) = 0$, will be added iteratively to the classifier.

The result is that the samples added to the system during learning are given by incorrect classification during system *life*. A new interactive session will provide additional exemplars to the training set if they were incorrectly classified. Therefore, the classifier evolves according to its perceptual experience. This focus is well suited for contexts like identity where the individual facial appearance changes in time, a fact which could not be completely tackled by a fixed training set.

4. Experiments

4.1. Video Streams Dataset

In the video streams context, a main problem is the absence of standard video stream databases with the complexity typical of HCI environments. Most facial databases do not contain sequences offering the facial evolution of different individuals. The availability of a controlled illumination and restricted background database such as XM2VTS [17] is not well suited to verify the unrestricted problem tackled in this paper.

Due to that reason, the data set used to carry out the experiments presented in this document contains different video streams that have been acquired and recorded using different standard webcams. The dataset contains 500 different video streams which correspond to approximately 430 individuals (250 males 180 females). These sequences were taken on different days without special illumination restrictions. Therefore, some were taken with natural (therefore, variable) and others with artificial illumination. The sequences cover different gender, face sizes and hair styles. They were taken at 15 Hz during $15-30$ seconds, i.e., each sequence contains from 210 to 450 frames of 320×240 pixels. All the frames contain at least one individual in unrestricted pose, i.e., there is a face in each frame but not always frontal. In the experiments considered, only the most salient frontal face was analyzed. Among the sequences acquired, only 70 of them correspond to *revisitors*.

4.2. Results

To define the PCA space, an independent dataset of 4000 still images different was used. This PCA space is fixed and not modified. Every exemplar is projected to that space to get its representation which is later classified by the n-class classifier. The weighted combination of the classifications provided for a meeting reports a suggested class. That given class is used to finally verify the identity assigned to the individual met.

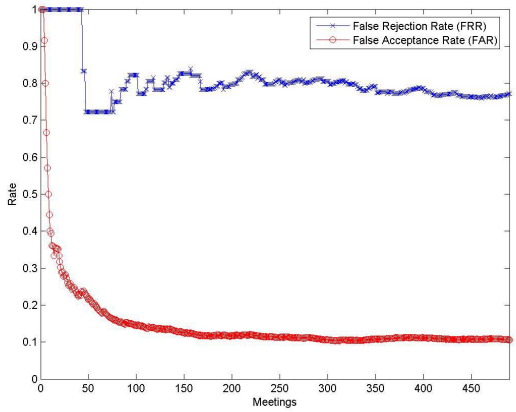

Figure 4. Results achieved for meetings with unknown individuals (FAR), 86% of the total number of meetings, and for already known individuals (FRR), 14% of the total number of meetings, along the system evolution.

Figure 4 shows the results achieved along the system evolution, starting from a *tabula rasa* perspective. These results have been averaged after 5 randomly ordered runs. For a specific meeting, the False Acceptance Rate (FAR) indicates the ratio, up to that moment, of the total number of meetings corresponding to unknown individuals which have been falsely accepted as known individuals. At the beginning the system seems to not have enough samples to model the unknown class, for that reason the error decreases notoriously until approximately 100 meetings, moment in which the error is lower than 15%.

On the other side, the False Rejection Rate (FRR) represents the ratio which corresponds to an already met identity which was falsely considered as unknown. This rate results are not good enough, approximately 80% of the identities are incorrectly not recognized, but we will try to examine the details. The total number of identities modelled are 432, corresponding only 10% of them, 43, to identities that the system met more than once. Under our assumption a familiar identity needs to store multiple patterns, we plotted in Figure 5 the evolution of the FRR for those identities which were met by the system more than three times. The number of meetings is still reduced, but for all the identities the error starts to be reduced (identities 46, 362 and 412) after the fourth meeting, or the FRR evolution improves and/or become stable, always clearly better than the average shown in Figure 4. Therefore, these preliminary results suggest that successive meetings with an identity serve to improve the identity model, i.e., to reduce its False Rejection Rate (FRR).

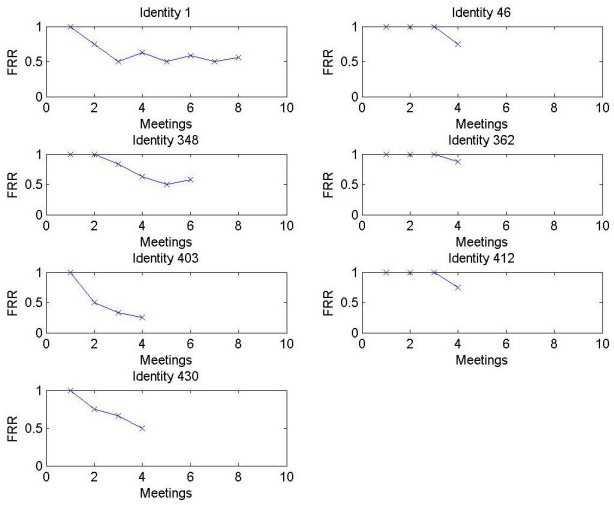

Figure 5. FRR evolution for identities with more than three meetings (45% of the total number of revisits).

5. Conclusions

As we mentioned above, our main objective is to recognize a reduced number of familiar identities in not necessarily high resolution images. At the same time the system must be able to reject individuals not belonging to the familiar group. These abilities must be reached using an automatic face detector and using no more facial data than those extracted automatically from the system meetings.

In the experiments carried out, the system shows an improving performance in terms of rejecting unknown individuals. However, the performance achieved for revisiting people is still far from being useful. Observing in detail the recognition rate evolution for repeated identities, it is observed that it tends to become better than the average. Can we assume that an individual becomes familiar when a collection of multi views is obtained? Future work must focus on collecting more meetings for familiar individuals in order to verify this hypothesis.

Future work should also consider the coordination with other modalities which could supervise the system in case of doubt. Note that, even if we try to recognize only a low number of individuals, face recognition may still fail because there is always the possibility that unseen facial images confuse the system (i.e. an individual in certain pose or under certain illumination is misrecognized).

Acknowledgments

Work partially funded by research projects Univ. of Las Palmas de Gran Canaria 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] A. Adler and J. Maclean. Performance comparison of human and automatic face recognition. In *Biometrics Symposium*, 2004.

[2] L. Aryananda. Recognizing and remembering individuals: Online and unsupervised face recognition for humanoid robot. In *Proc. of IROS*, 2002.

[3] E. Bailly-Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, and J.-P. Thiran. The banca database and evaluation protocol. In J. Kittler and M. Nixon, editors, *Proc. Audio- and Video-Based Biometric Person Authentication*, pages 625–638, Berlin, June 2003. Springer.

[4] M. Bartlett and T. Sejnowski. Independent component of face images: a representation for face recognition. In *Procs. of the Annual Joint Symposium on Neural Computation,Pasadena, CA*, May 1997.

[5] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. on PAMI*, 19(7):711–720, 1997.

[6] A. M. Burton, R. Jenkins, P. J. Hancock, and D. White. Robust representations for face recognition: Te power of averages. *Cognitive Psychology*, 51(3):256–284, 2005.

[7] A. M. Burton, S. Wilson, M. Cowan, and V. Bruce. Face recognition in poor quality video: Evidence from security surveillance. *Psychological Science*, 10(3), 1999.

[8] M. Castrillón Santana, O. Déniz Suárez, M. Hernández Tejera, and C. Guerra Artal. Real-time detection of faces in video streams. In *2nd Workshop on Face Processing in Video*, Victoria, Canada, May 2005.

[9] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings IEEE*, 83(5):705–740, 1995.

[10] 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.

[11] J. Fan, N. Dimitrova, and V. Philomin. Online face recognition system for videos based on the modified probabilistic neural networks. In *Proceeding of IEEE ICIP 2004*, pages 104–110, Singapore, 2004.

[12] C. Guerra Artal. *Contribuciones al seguimiento visual precategórico*. PhD thesis, Universidad de Las Palmas de Gran Canaria, Octubre 2002.

[13] R. Kemp, N. Towell, and P. G. When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, 11(3):211–222, 1997.

[14] 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.

[15] V. Krueger and S. Zhou. Exemplar-based face recognition from video. In *European Conference on Computer Vision (ECCV), Kbenhavn, Denmark*, May 2002.

[16] H. Kruppa, M. Castrillón Santana, and B. Schiele. Fast and robust face finding via local context. In *Joint IEEE Internacional Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 157–164, October 2003.

[17] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, March 1999.

[18] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 107–119, January 2000.

[19] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE Conference on Computer Vision and Pattern Recognition 2005*, 2005.

[20] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face recognition algorithms. TR 6264, NISTIR, January 1999.

[21] G. Pike, R. Kemp, and N. Brace. The psychology of human face recognition. In *IEEE Colloquium on Visual Biometrics*, 2000.

[22] A. Tolba, A. El-Baz, and A. El-Hardy. Face recognition: A literature survey. *International Journal of Signal Processing*, 2(1):88–103, August 2005.

[23] L. Torres and J. Vilá. Automatic face recognition for video indexing appplications. *Pattern Recognition*, 35:615–625, March 2002.

[24] M. Turk. Computer vision in the interface. *Communications of the ACM*, 47(1):61–67, January 2004.

[25] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

[26] V. Vapnik. *The nature of statistical learning theory*. Springer, New York, 1995.

[27] P. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[28] G. Wallis and H. Buelthoff. Learning to recognize objects. *Trends in Cognitive Sciences*, 3(1):22–31, January 2000.

[29] C. Wallraven and H. Buelthoff. Automatic acquisition of exemplar-based representations for recognition from images sequences. In *Computer Vision and Pattern Recognition*, 2001.

[30] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM*, 35(4):399–458, 2003.