Proceedings of IEEE Workshop on Content-Based Access of Images and Video Libraries, pp. 55-59, 2000

Semantic Access of Frontal Face Images: the expression-invariant problem

Aleix M. Martínez
Robot Vision Lab, School of Electrical and Computer Engineering, Purdue University, IN 47906
Sony Computer Science Lab, 6, rue Amyot, 75005 Paris
aleix@ecn.purdue.edu

Abstract

Semantic queries to a database of images are more desirable than low-level feature queries, because they facilitate the user's task. One such approach is the object-related image retrieval. In the context of face images, it is of interest to retrieve images based on people's names and facial expressions. However, when images of the database are allowed to appear at different facial expressions, the face recognition approach encounters the expression-invariant problem, i.e. how to robustly identify a person's face for which its learning and testing face images differ in facial expression. *This paper presents a new local, probabilistic approach that accounts for this (as well as other previous studied) difficulty.*

1 Introduction

The rapidly increasing interest in content based image retrieval systems is fueled by the possibilities that are envisioned in many diverse domains. One such fertile domain consists of retrieving face images from large multimedia databases, as might be produced by television and newspaper establishments. In this paper, we present an approach that allows to obtain frontal face images by means of semantic queries, such as "images of Bill smiling". Semantic queries are more desirable than low-level feature queries, because they facilitate the user's task [15]. For illustration, we show the case of Figure 1, where different images of the same subject require different queries.

The above stated scheme implies that face images of different subjects may appear at different facial expressions, illumination conditions and/or under certain occlusions; which makes the problem difficult to be resolved. However, although much efforts have been made to solve the illumination problem (e.g. [1, 2]) and some to resolve the occlusion problem (e.g. [16]), little has been done to account for the recognition of faces that appear at different facial expressions; i.e. expression-invariant recognition. The problem of expression-invariant recognition can be formulated as follows: *"how can we robustly identify a person's face for which its learning and testing mug images differ in facial expression?"*. An example is shown in Figure 1, where image (a) is used for learning and images (b-d) for recognition.

In this communication we present a system that partially compensates for this difficulty. This we will do within the eigenspace domain [12, 21, 1, 18]. The eigenspace approach has proven to be very useful (with high recognition rates) for the recognition of human faces, even where the facial expression of the learning and testing images do not differ too much [10]. However, this approach would by no means be able to handle the case of Figure 1(d). This is obvious not only in practice, but also theoretically [6, 22]: Since the eigenspace representation can be seen as an invariant paradigm (where the second-order statistics of the image set represent those features that are expected to be invariant to the above mentioned image/object variations), it follows that there is always a set of images for which the learned measure (or function) will not be optimal (see for example [22] p.20 and references therein). In a recent paper [17], Martínez and Kak have also shown that the switch from "simple" pattern recognition techniques (e.g. PCA) to more "discriminant" techniques (e.g. LDA – Linear Discriminant Analysis) is not always warranted and may sometimes lead to faulty systems design, where small (or non-representative) training data-sets are used.

In this contribution, we show that knowing the facial expression of a face can help us to overcome part of the difficulties of expression-invariant recognition. This we will found in two known facts. *(i)* It is now known that different facial expressions affect certain parts of the mug face more than others [20, 5, 11, 19]; thus knowing the facial expression of face images will help us to determine which areas of the mug face are best to be analyzed. Furthermore, this point can be easily adapted to our local probabilistic approach [16], as we shall discuss later. *(ii)* From Ekman and Friesen's work [9], we know which facial muscles play a role in each facial expression, which can help us to (partially) "erase" certain expressions from the face. While the first method is the main contribution of the paper, we show how the second (which is related to [3, 4]) can be

Figure 1. (a) Learning sample. (b-d) Testing (query) images.

used as an extension (or improvement) of the first.

2 Semantic retrieval of face images

In [15], we introduced an approach to automatically index a database of images by object categories, which ultimately allowed "semantic" access to the database based on these object names; i.e. what is known as object-related image retrieval. In this paper, we extend this idea to the case of human-face class (category). We will allow, not only access to this category, but also identity (people's name) and facial expression.

Localization of the face, eyes and mouth is done using an eigenspace representation as in [18]. The boundaries of the mug face are localized by means of a contour technique. And, finally, the mug face is wrap to a final "standard" shape. After wrapping all faces have the eyes, mouth and boundaries of the face to roughly the same position. To identify faces, we use a local probabilistic PCA approach with similar underlying ideas as those described in [16]. Facial expressions are analyzed using a Gabor jet approach, which has already proven to do a good job on the task [8, 13]. The rest of this section sketches the identification approach, the recognition of facial expressions and some experimental results.

2.1 Local probabilistic approach

One way to deal with partially occluded objects (such as faces) is by using local approaches [18, 16]. In such techniques, a face is divided into several parts that are analyzed in isolation, and then a voting space is used to search for the best match. However, a voting technique can easily miss-classify a test image, because it does not take into account how good a local match is (in relation to other local matches). In this paper, we summarize our probabilistic approach and extend it to solve the expression-invariant problem (for results on partially occluded faces the reader is referred to [16]).

Learning stage: To learn a set of face images, we will use an eigenspace representation [12, 21, 1, 18]. To generate the local eigenspaces, we first localization and wrap the mug part of the face to a "standard" (general) shape. We then generate the learning set

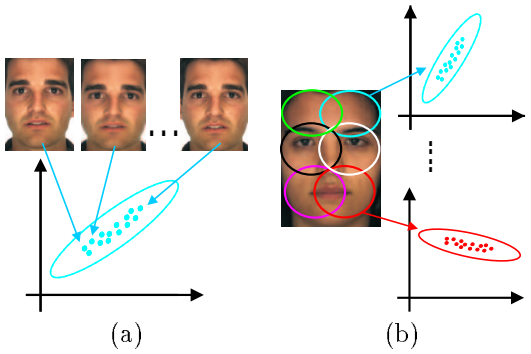


Figure 2. (a) We learn the subspace where different localization lie. (b) The probabilistic local approach

$\mathbf{X}_k = \{\mathbf{x}_{1,1,k}, \ldots, \mathbf{x}_{n,m,k}\}$; where $\mathbf{x}_{i,j,k}$ is the k^{th} local area of the j^{th} (wrap) sample image of class i (in its vector form), n is the total number of classes (i.e. people), and m the number of samples per class. In this contribution, we assume that the number of samples is equal for all classes (in any other case, derivations might slightly differ from the ones described below). We also assume $k = \{1, 2, 3, 4, 5, 6\}$, i.e. six local areas. To obtain each $\mathbf{x}_{i,j,k}$ sample an ellipse-shape area $x^2/d_x^2 + y^2/d_y^2 = 1$ is used. After finding the mean feature vector μ_k for every $\mathbf{X}_k$, the covariance matrices $\mathbf{Q}_k = (\mathbf{X}_k - \mu_k)(\mathbf{X}_k - \mu_k)^T$, $\forall k$ are computed. Eigenspaces are obtained by taking the p eigenvectors associated with the largest eigenvalues of $\mathbf{Q}_k$. In the sequel, we shall refer to these eigenspaces as $\mathcal{E}ig_k$, and to their projecting matrices as $\mathbf{E}_k$.

However, the eigenspace representation obtained above does not account for the localization problem. The localization error problem is encountered when a *small* localization error makes the eigen-representation of a test image close to an incorrect class. This point becomes critical when the learning and testing images differ on facial expression and this is why a method to model such a variation is needed [16]. To resolve this problem, we need to project all learning mug faces (accounting for the localization error) onto the above computed eigen-representation; Figure 2. Mathematically speaking, we define $\hat{\mathbf{X}}_k = \{\hat{\mathbf{x}}_{1,1,k}, \ldots, \hat{\mathbf{x}}_{n,m,k}\}$, where $\hat{\mathbf{x}}_{i,j,k} = \{\mathbf{x}^1_{i,j,k}, \ldots, \mathbf{x}^r_{i,j,k}\}$ and represents all possible images accounting for all r possible errors of localization, where r is of the order of 2^{2f}, being f the number of features (e.g. eyes, nose, etc.) localized in the face. In our previous work, we have shown that this localization error has a variance of ± 3 by ± 4 pixels when using the above mentioned localization method on the images of the AR database of faces [16]. For obvious reasons, each $\hat{\mathbf{x}}_{i,j,k}$ set is only projected onto its corresponding eigenspace by means of $\mathbf{E}_k$. Each $\hat{\mathbf{x}}_{i,j,k}$ set is expected to be within a small subspace of its cor-

responding $\mathcal{E}ig_k$, which can be modeled by means of a Gaussian distribution $\mathcal{G}_{i,j,k}$ with an associated mean $\mu_{i,j,k}$ and covariance matrix $\Sigma_{i,j,k}$; where $\mathcal{G}_{i,j,k}$ is the Gaussian model associated with the training sample $\hat{\mathbf{x}}_{i,j,k}$, and $\Sigma_{i,j,k} = (\hat{\mathbf{x}}_{i,j,k} - \mu_{i,j,k})(\hat{\mathbf{x}}_{i,j,k} - \mu_{i,j,k})^T$. This is depicted in Figure 2. Notice that the set $\hat{\mathbf{x}}_{i,j,k}$ (which can be very large) needs not be stored in memory, only the Gaussian model (mean and covariance matrix) is needed for consecutive computations.

Identification stage: When a test image $\mathbf{t}_z$ is to be recognized, we work as follows. We first localize the face and wrap it to a 120 by 170 pixel array (which we shall denote as $\bar{\mathbf{t}}_z$). We then project each of the six local areas onto the above computed eigenspaces. More formally, $\hat{\mathbf{t}}_{z,k} = \mathbf{E}_k \cdot \bar{\mathbf{t}}_{z,k}$, $\forall k$; where $\bar{\mathbf{t}}_{z,k}$ represents the k^{th} local area of $\bar{\mathbf{t}}_z$, and $\hat{\mathbf{t}}_{z,k}$ its projection onto $\mathcal{E}ig_k$. Since the mean feature vector and the covariance matrix of each local subspace are already known, the probability of a given match can be directly associated with a suitably defined distance $(\hat{\mathbf{t}}_{z,j,k} - \mu_{i,j,k})\Sigma_{i,j,k}(\hat{\mathbf{t}}_{z,j,k} - \mu_{i,j,k})$, i.e. the Mahalanobis distance. Mathematically, $LocRes_{i,k} = \sum_{j=1}^{m} Mh(\hat{\mathbf{t}}_{z,k}, \mathcal{G}_{i,j,k})$; where $Mh(\cdot)$ is the Mahalanobis distance, and $LocRes_{i,k}$ the recognition result of the k^{th} local area of class i.

Finally, we add all local distances (probabilities), $Res_i = \sum_{k=1}^{6} LocRes_{i,k}$, and search for the minima (maxima), $RecClass = argmin_i\, Res_i$; where $ResClass \in [1, n]$. If a video sequence is supplied, we keep adding distances (probabilities) for each of the images and only compute the minima (maxima) at the end of the sequence or when a threshold has been reached.

2.2 Facial expression recognition

In order to detect facial expressions in pre-localized (wrap) face images, a Gabor jet approach is used; similarly as in [8]. Gabor filters, which have been extensively used in computer vision, are obtained by modeling a sine wave with a Gaussian envelope. Such filters are expected to remove most of the variability in images due to lighting and contrast changes, and closely model the response of several visual cortical cells [7].

Mathematically, every image $I(\mathbf{x})$ is convoluted with a plane wave (Gabor) kernel

$\psi(\mathbf{k}, \mathbf{x}) = \frac{\mathbf{k}^2}{\sigma^2} e^{-\mathbf{k}^2\mathbf{x}^2/2\sigma^2} \left[e^{i\mathbf{k}\,\mathbf{x} - exp\frac{-\sigma^2}{2}} \right]$

where ψ results in a planed wave curved characterized by the vector $\mathbf{k}$ enveloped by a Gaussian function with parameter σ (which determines the window). This representation allows description of spatial frequency structure in the image while preserving information about spatial relations. The vector k is defined as

$\mathbf{k} = \begin{pmatrix} f_v cos\phi_u \\ f_v sin\phi_u \end{pmatrix}$, $where\ f_v = 2^{-\frac{v+2}{2}} and\ \phi_u = u\frac{\pi}{8}$

v and u define the frequency and orientation of the kernels respectively. In our experimental results, v takes values from 0 to 4 and u from 1 to 8.

In order to identify facial expressions in face images, we work as follows (similarly as in [8]). First 50 face images (of 50 different individuals) with a neutral expression are wrap to the "standard" shape described earlier. Second, an average image M is obtained by averaging the graylevel value of each pixel location. Third, we compute the image $\check{\mathbf{t}}_z = \bar{\mathbf{t}}_z - M$ for every new incoming wrap face image $\bar{\mathbf{t}}_z$. After downsizing these images to a 30 by 42 pixel array, the Gabor filter images are obtained. Identification of facial expression is obtained by performing nearest neighbor classification.

2.3 Experimental results

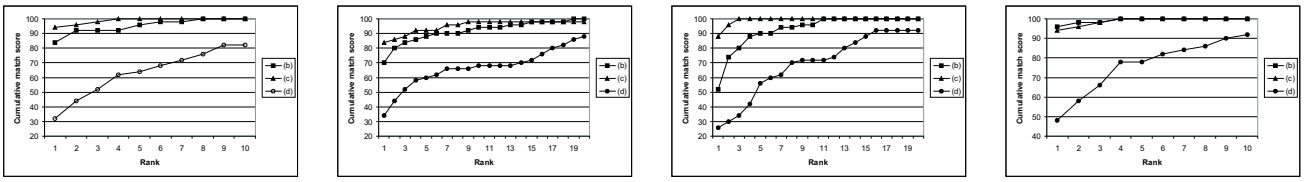
In order to assess how good the above described system is to retrieve images from a database of frontal faces, we applied both methods, identification (naming) and classification of facial expressions, to images of 50 subjects (25 females and 25 males) of the AR-face database [14].[1] Four images per subject were used; each of which describes a distinct facial expression: neutral, smiling, anger and "screaming". For illustration, these images for one subject are shown in Figure 1. The neutral (without expression) images were used for learning, while the others for identification. Facial expressions were correctly classified in 90% of the images. Results of identification (for each group of distinct facial expression) for $p = 20$ are displayed in Figure 3(a) as a function of rank (x-coordinates) and successive match score (y-coordinates) as described in [10]; i.e. rank R states that the correct class is within the R first choices. Note that whereas this has no much importance within certain application of face recognition (e.g. security access), it does indeed correspond to the image retrieval scenario.

3 Expression-invariant Recognition

From the above experimental results (and as intuition would suggest) it seems clear that different facial expressions lead to different recognition success rates. For example, it is interesting to see that those images expressing anger are much easy to be identified (named) than those ones expressing happiness; Figure 3. In fact, this should not be very much surprising. When one pays attention to both types of images (anger and happiness), it appears evident that those expressing happiness diverge more from the neutral image (the one used for learning purposes) than what those expressing anger do.

Obviously, one way to cope with this difficulty consist in morphing all testing images to equal (in shape)

[1] The AR database of face images is publicly available at `http://RVL.www.ecn.purdue.edu/~aleix/`



(a) (b) (c) (d)

Figure 3. Experimental results.

the learning(s) one(s) [3]. Unfortunately, this cannot always be achieved; e.g. the eye's area of a "screaming" image (Figure 1(d)) cannot be morphed to a neutral eyes expression, because in most of the faces the texture of the inside of the eyes is not available (and can neither be inferred from the training data [3]).

Other (psychological) evidences can also be of some help. It was proposed by Sackheim [20] that emotions (i.e. facial expressions) are most expressed on the left side of the face. If such statement was true, our recognition system should best analyze the right side of the face (because this is the one that diverges less from the neutral face image). Critics of this type of experiments have noted that such studies fail to distinguish between photograph of posse emotions and those of genuine emotions [11]. This asymmetry seems to be less evident for spontaneous, genuine emotions. New experiments seem to come up with a better explanation. Since anger and sadness involve the right hemisphere more than the left, while happiness involve the left, it is logical to assume that anger will be more express on the left side and happiness on the right. [5] showed that composite photos constructed of the left side of the face are judged as happier, and composites of the right side of the face are judged as sadder.

In order to analyze either the above statement can be of some help or not, a totally independent experiment was computed. This consists of identifying faces based only on the left or on the right local areas among the six local parts described earlier, which obviously gives us two cumulative-match-score/rank curves per facial expression group, Figure 3(b-c). It is easy to see that for those face images expressing: (b) happiness, the left side of the face gives better results than the right, (c) anger, the right side achieve better recognition results than the left, (d) complex expressions (e.g. "screaming") do not have a clear preferred side.

It is also well-known that different facial expressions influence different parts of the face more than others [9]. For example, while anger is (in general) more emphasized on the top of the face (e.g. with vertical lines appearing there) and the mouth area, happiness is also very emphasized on the eye's area. To analyze this effect, a new test was considered. In this case only one local part was used for recognition purposes, which leads us with six different recognition curve per facial expression. These results clearly suggest that different areas (left eye area, right eye area, etc.) contribute differently to recognition depending on the facial expression being expressed.

Mathematically, we can esily model this effect: $LocRes_{i,k} = \sum_{j=1}^{m} w_{j,k} Mh(\hat{\mathbf{t}}_{z,k}, \mathcal{G}_{i,j,k})$, where $w_{j,k}$ are the weights associated to each of the six local areas $k \in [1,6]$ given a facial expression $j \in [1,m]$. These weights are to be learned from some training data. Notice that this implies the set of learning data to be labeled (not only with identity, but also with facial expression). The weights can be easily determined $w_{j,k} = q_{j,k} / \sum_{k=1}^{6} q_{j,k}$, where $q_{j,k}$ states for the number of images with expression j successfully identified using only the k^{th} local area. It is important that the reader keeps in mind that the data used for training these weights (for example the one used earlier in the experimental results of section 2) can never be the same used for testing purposes. Notice thus that learning is now divided in two different stages. In the first one, the eigenspaces and the Gaussian-like subspaces (which accounts for the localization error) are learned. In the second, the local weights $w_{j,k}$ are settled. The final tests (retrieval) must be computed on a totally independent set of images.

3.1 Experimental results

In this section, the same data as in section 2.3 is used for learning purposes (i.e. the neutral face images are used to learn the eigen-representations and the others to compute the local weights $w_{j,k}$). Testing was done using a totally independent group of 50 different people (25 females and 25 males), randomly selected from the rest of the images of the AR-face database. For each subject only images expressing happiness, anger and "screaming" were considered. Recognition results as a function of cumulative match score and rank are displayed in Figure 3(d). Notice how recognition of faces is more expression invariant now.

3.2 Standard shape

From the seminal work of Ekman and colleges (e.g. [9]), it is known which areas of the face are most expected to change and how. This information can be learned using an "optical flow" method as follows.

First a 2D deformation representing a change in expression (e.g. a testing image expressing happiness and a neutral expression prototipe—such as M) is estimated from two (wrap) face images; i.e. optical flow between both images. Next, this flow is mapped onto the new face expressing happiness that wants to be morphed to a neutral expression [3, 4]. However, instead of morphing the entire mug image each time, we will only morphed those parts that are supposed to be affected. This can be known form both [9] and/or our learned weights $w_{j,k}$. Obviously, this extension cannot be applied to "open" the shot eyes of Figure 1(d), because the texture that describes them is not known and cannot be inferred. In those cases, alternatives, such as morphing the learning image to equals the testing, might be of interest.

As a simple example to show the possibilities of such an extension, this optical flow method was used to morphed the mouth area of those images expressing happiness and anger to a neutral expression (neutral mouth of image M). Results went up to 100% for $R = 3$.

4 Conclusions

This paper describes a system able to retrieve images from a database that contains frontal view faces of different individuals. These images are allowed to diverge in expression, which arises the expression-invariant problem. To solve this, we have proposed a new approach based on how expressions effect different areas of the face [9, 20, 5, 11, 19]. Then, we have used of a morphing stage to make the testing and learning images equals in shape as an extension of the first approach. This was done using an optical flow approach as in [3, 4]. Experimental results show that images cannot only be better retrieved using this method, but also that different facial expressions can obtain similar identification success rates, i.e. retrieving results can be close to expression-invariant.

References

[1] P.N. Belhumeour, J.P. Hespanha and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," IEEE **PAMI**-19(7):711-720, 97.

[2] P.N. Belhumeour and D.J. Kriegman, "What is the Set of Images of an Object Under All Possible Lighting Conditions," Int. J. Comp. Vision, 98.

[3] D. Beymer and T. Poggio, "Face Recognition from One Example View," Science 272(5250), 96.

[4] M.J. Black, D.J. Fleet, Y. Yacoob, "A Framework for Modeling Appearance Change in Image Sequences," Proc. ICCV, pp. 660-667, 98.

[5] R. Campbell, "The leteralization of emotion: A critical review," International J. of Psychology 17:211-219, 82.

[6] D.J. Clemens and D.W. Jacobs, "Space and time bounds on indexing 3-D model from 2-D images," IEEE Trans. **PAMI**-13(10):1007-1017, 90.

[7] J.G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," J. Optical Society of America A 2:1160-1169, 65.

[8] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman and T.J. Sejnowski, "Classifying Facial Actions," IEEE Trans. **PAMI**-21(10):974-989, 99.

[9] P. Ekman and W. Friesen, "Facial Action Coding System: A technique for the measurements of facial movements," Consulting Psychologists Press, 78.

[10] P.J. Phillips, H. Moon, P Rauss and S.A. Rizvi, "The FERET evaluation methodology for face-recognition algorithms," Proc. AVBPA, 97.

[11] J. Hager, "Asymmetries in facial expression," In P.Ekman Ed. *Emotion in the human face*, pp. 318-352, Cambridge University press, 82.

[12] M. Kirby and L. Sirovich, "Application of the Karhunen-Love procedure for the characterization of human faces," IEEE Trans. **PAMI**-12(1):103-108, 90.

[13] M.J. Lyons, J. Budynek and S. Akamatsu, "Automatic Classification of Single Facial Images," IEEE Trans. **PAMI**-21(12):1357-1362, 99.

[14] A.M. Martínez and R. Benavente, "The AR face database," CVC Tech. Rep. #24, June 98.

[15] A.M. Martínez and J.R. Serra, "Semantic Access to a Database of Images: AN approach to object-related image retrieval," Proc. of IEEE Multimedia Computing and Systems, June 99.

[16] A.M. Martínez, "Recognition of Partially Occluded and/or Imprecisely Localize Faces Using a Probabilistic Approach," Proc. IEEE CVPR'2000, June 2000.

[17] A.M. Martínez and A.C. Kak, "PCA versus LDA," IEEE Trans. **PAMI**, 2000.

[18] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," IEEE Trans. **PAMI**-19(7):696-710, 97.

[19] M. Moscovitch and J. Olds, "Asymmetries in spontaneous facial expressions and their possible relation to their hemispheric specialization," Neuropsychologia 20:71-81, 82.

[20] H. Sackheim, R.C. Gur and M.C. Saucy, "Emotions are expressed more intensively on the left side of the face," Science 202:434-436, 78.

[21] M. Turk and A. Pentland, "Eigenfaces for recognition," Journal Cognitive Neuroscience 3(1):71-86, 91.

[22] S. Ullman, "High-level Vision: Object recognition and visual cognition," MIT press, 1996.