ABSTRACT

In this article we present a multimodal system to person recognition by integrating two complementary approaches that work with video data. The first module exploits the behavioural information: it is based on statistical features computed using the displacement signals of a head; the second one is dealing with the physiological information: it is a probabilistic extension of the classic Eigenface approach. For a consistent fusion, both systems share the same probabilistic classification framework: a Gaussian Mixture Model (GMM) approximation and a Bayesian classifier. We assess the performances of the multimodal system by implementing two fusion strategies and we analyse their evolution in presence of artificial noise.

Index Terms— Identification of persons, Face recognition, Object recognition.

1. INTRODUCTION

For decades human face recognition has been an active topic in the field of object recognition. Most of algorithms have been proposed to deal with individual images, also called image-based recognition, where both the training and test sets consist of individual face images. However, with existing approaches, the performance of face recognition is affected by different kinds of variations, for example: expression, illumination and pose changes. Thus, researchers have started to look at video-based recognition, in which both training and test sets are video sequences containing the face. A detailed analysis of person recognition using still images, its performances and limitations can be found in [1, 2].

Person recognition using videos has some advantages over image-based recognition. First, the temporal information of faces can be exploited to facilitate the recognition task; for example, dynamical characteristics, which are specific to each person, like the motion of the head, the evolution of the pose or the mimic of the face. Second, more effective representations, such as 3D face models or super resolution images, can be obtained from the video sequence and used to improve the performance of the systems. Finally, video-based recognition enable learning or updating the subject model over time.

It's a common trend in literature to exploit only a part of the video information; in fact in our research experience [2, 3] and in the majority of research studies the recognition systems have been based either on the physiological information (facial appearance) either on the temporal one (facial motion). Considering the potential of these two independent modalities, the natural evolution of the video-based person recognition is directed towards the study of a multimodal recognition system, that is exploiting all the video information; consequently, we intend to investigate this original perspective and make use of the complementary nature of these two modalities, in order to develop a system with improved discriminating power and more robustness.

In this paper, we present a multimodal person recognition system, which is composed by two complementary modules. The first one [3], is based on displacement signals of a few head features, automatically extracted from the video sequence; statistical features are then computed from these signals and used for discriminating identities. The second system is a probabilistic extension of the classic Eigenface approach [4], in which the recognition task is done on a reduced face space, computed by using a Principal Component Analysis (PCA) transformation. For a consistent fusion, both systems share the same probabilistic classification framework: the characteristic head displacements and the personal variations in face space are modelled by Gaussian Mixture Models (GMM) and the classification task is solved as a Bayesian decision making problem. The experimental results show that the multimodal integration provides an important advantage in discriminating identities, especially in presence of corrupting noise.

The main contribution of this article is the development of a multimodal framework that operates a fusion of two complementary video modalities: a well established physiolog-
ical one, related to facial appearance, and a pioneering beha-

vioral one, related to head motion. Moreover there are

some important side contributions as: 1) the analysis on noise

robustness of the two unimodal systems and the multimodal

one proposed; 2) the analysis on the effectiveness of unbal-

anced fusion (the fusion between a performing and a weak

system), which is still an open question in literature.

The rest of the paper is organized as follows: in section 2

we detail our recognition system, then we report and comment

experiments in section 3, and finally we conclude this paper

with remarks and future work in section 4.

2. DESCRIPTION OF THE RECOGNITION SYSTEM

Our person recognition system can be organized in three
different modules: a Static Recognizer, which computes its
recognition scores by using the facial appearance of the sub-
ject, a Temporal Recognizer, which computes its scores by ex-
ploring the facial motion of the individual, and a Fusion Mod-
ule, which achieves the identification and verification task by
integrating the two previous modalities.

2.1. Static Recognizer module

Our appearance-based recognition algorithm is a probabilistic
extension of the classic Eigenface approach, presented in [4]
by Turk and Pentland.

Following the original technique, we compute the Prin-
cipal Component Analysis (PCA) on a general set of face im-
ages, in order to obtain a set of orthogonal vectors (the eigen-
vectors) that optimally represent the distribution of the data
in the Root Mean Squares (RMS) sense; these vectors define
the sub-space of face images, which we call face space. A
new face image is transformed into its eigenface components
by a simple projection into the face space; we will denote the
projected vector for image \( k \) as: \( y_k \).

Then, we improve the original identification and verifica-
tion task by using a Bayesian framework. Firstly, for each
individual we want to model the distribution of his images in
the face space; we approximate the class-conditional proba-
bility density function of the individual by using a Gaussian
Mixture Model (GMM). We can express using the following
formula:

\[
P (y_k | \omega_q) = \sum_{c=1}^{C} \alpha_c \mathcal{N} (y_k; \mu_c, \Sigma_c) \tag{1}
\]

where \( \omega_q \) refers to class (individual) \( q \), \( \alpha_c \) is the weight of
the \( c \)-th Gaussian component, \( \mathcal{N} \). After that, in order to test
a given image, we compute the log-posterior probabilities for
each class \( q \) (we will refer to them as \( \beta \)):

\[
\beta_{q,k} = \log (P (\omega_q | y_k)) = \log \left( \frac{P (y_k | \omega_q) P (\omega_q)}{P (y_k)} \right) \tag{2}
\]

For completeness, we note that the priors and scaling factors,
presented in the previous formula, are directly estimated from
the training database.

2.2. Temporal Recognizer module

A detailed description of this module can be found in our pre-
vious publication [3]; here we provide a short summary of the
algorithm.

Head motion is firstly analysed by retrieving the displace-
ments of the eyes, nose and mouth in each video frame. Then,
the raw signals are transformed and normalized in order to ob-
tain video independent feature vectors, in a way that the scale
and geometrical parameters do not interfere with our recog-
nition results. Finally, we model the distribution of character-
istic displacements (embedded in feature vectors) over time
by training individual Gaussian Mixture Models (GMM), and
we achieve classification through a Bayesian classifier.

For integrating this module in our multimodal system, we
recover all log-posterior probabilities (the similarity scores)
computed during the Bayesian classification (we will refer to
them as \( \gamma \)):

\[
\gamma_{q,k} = \log (P (\omega_q | X_k)) \tag{3}
\]

in which \( \omega_q \) refers to class \( q \) and \( X_k \) is a matrix containing
the behavioural features extracted from video \( k \).

2.3. Fusion Module

The Fusion Module integrates the two similarity scores (log-
posterior probabilities) from the previous modules and com-
putes identification and verification rates of the multimodal
system. In this paper, the multimodal similarity score is cal-
culated by two versions of the weighted sum fusion rule [5],
which in our case has the following general formula:

\[
\theta_{q,k} = b_{q,k} \beta_{q,k} + g_{q,k} \gamma_{q,k} \tag{4}
\]

in which \( b_{q,k} \) and \( g_{q,k} \) are the two weights.

In the first implementation, we compute the mean be-
tween the scores of the two modules (equal weighting):

\[
b_{q,k} = g_{q,k} = 0.5 \quad \forall q, k \tag{5}
\]

This simple technique has an interesting probabilistic inter-
pretation. If we assume that \( X_k \) and \( y_k \) are independent from
each other and equally distributed, and that all the classes are
equiprobable (a common scenario in real applications), then
the similarity score \( \theta_{q,k} \) is the joint log-posterior probability
of \( X_k \) and \( y_k \):

\[
\theta_{q,k} = 0.5 \log (P (\omega_q | y_k, X_k)) + T \tag{6}
\]

where \( T \) is a constant translating factor.
3. EXPERIMENTS AND RESULTS

3.1. Video database

Unfortunately, existing standard video databases do not match the requirements for efficiently testing our algorithms; in particular, the Temporal Recognizer module needs a few minutes for each individual, in order to extract the temporal information and train the GMM models. For this reason, we collected a set of 192 video sequences of 12 different persons, for the task of training and testing our system. The video chunks are showing TV speakers, announcing the news of the day: they have been extracted from different clips during a period of 14 months. A typical sequence has a spatial resolution of $352 \times 288$ pixels and a temporal resolution of $23.97$ frames/second, and lasts almost 14 seconds (refer to Figure 1 for an example). Even though the videos are of low quality, compressed at $300$ Klits/second (including audio), the behavioural approach of our system is less affected by the visual errors, introduced during the compression process, than the pixel-based methods. Moreover, the videos are taken from a real case: the behaviour of the speakers is natural, without any constraint imposed to their movement, pose or action.

Fig. 1. The first 4 frames of a video sequence.

The second implementation of the similarity score for the integrated system is an adaptive weighting, successfully applied by Chang et al. [6], and computed as follows:

\[
\begin{align*}
    b_{q,k} &= \frac{\beta_{q,k} - \beta_{q}}{\sigma_{q,k}^2 + \sigma_{q}^2} \\
    g_{q,k} &= \frac{\gamma_{q,k} - \gamma_{q}}{\sigma_{q,k}^2 + \sigma_{q}^2}
\end{align*}
\]

in which $\beta_{q,k}$ and $\gamma_{q,k}$ are the $i$-th best scores for a given test $k$.

The general idea of this weighting scheme is that, if the difference between the first and second scores is large compared to the typical one, then the modality can be considered reliable and its weight is big. In our implementation, we normalize the scores to sum to 1: $b_{q,k} + g_{q,k} = 1$ $\forall q,k$.

3.2. Image database

Concerning the Static Recognizer module, we created an image database derived from the video database depicted before; for each training video we extracted 28 frames (2 frames/second), while for the testing set we retrieved only the first frame. Due to the well known high sensitivity of PCA-based recognition algorithms to facial alignment, variation in pose and scale, we manually normalized the image database by cropping the face region, then aligning and (in-plane) horizontally rotating the heads.

3.3. Experimental set-up

In our experiments, for the Temporal Recognizer module we selected 96 sequences for training (8 for each of the 12 individuals), and the remaining 96 (out of 192) were left for testing. For a detailed discussion of the parameters of this module, please refer to [3].

Concerning the Static Recognizer module, we selected a total of 2688 images for training (224 per individual) and 96 for testing; though, the algorithm is actually working with 5376 training images (448 per individual) due to vertical mirroring of original images. In our experiments we were obliged to chose a small eigenspace of dimension 13, due to the difficulty of approximating high dimensional distributions with a limited amount of training data; for the same reason we also considered GMMs with $1 \div 3$ Gaussian components for each model.

3.4. Results

The recognition results of the two disjoint modules are summarized in Tables 1 and 2, while the scores for the multimodal system using the mean and the adaptive weighting fusion operators are presented in Tables 3 and 4, respectively. The seco-

<table>
<thead>
<tr>
<th>Method</th>
<th>1-Best (%)</th>
<th>3-Best (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR (0)</td>
<td>93.75</td>
<td>97.92</td>
<td>6.91</td>
</tr>
<tr>
<td>SR (1)</td>
<td>90.63</td>
<td>97.92</td>
<td>3.88</td>
</tr>
<tr>
<td>SR (2)</td>
<td>71.88</td>
<td>90.63</td>
<td>8.66</td>
</tr>
<tr>
<td>SR (3)</td>
<td>55.21</td>
<td>81.25</td>
<td>15.53</td>
</tr>
</tbody>
</table>

Table 1. Identification and verification results for the Static Recognizer (SR).

<table>
<thead>
<tr>
<th>Method</th>
<th>1-Best (%)</th>
<th>3-Best (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR (0)</td>
<td>92.71</td>
<td>97.92</td>
<td>6.91</td>
</tr>
<tr>
<td>TR (0.333)</td>
<td>84.38</td>
<td>97.92</td>
<td>7.10</td>
</tr>
<tr>
<td>TR (1.333)</td>
<td>68.75</td>
<td>91.67</td>
<td>13.87</td>
</tr>
<tr>
<td>TR (3.333)</td>
<td>63.54</td>
<td>84.38</td>
<td>18.37</td>
</tr>
</tbody>
</table>

Table 2. Identification and verification results for the Temporal Recognizer (TR).
Table 3. Identification and verification results obtained by fusioning with the mean operator.

<table>
<thead>
<tr>
<th>Method</th>
<th>1-Best (%)</th>
<th>3-Best (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR (0) + TR (0)</td>
<td>94.79</td>
<td>97.92</td>
<td>2.18</td>
</tr>
<tr>
<td>SR (1) + TR (0.333)</td>
<td>92.71</td>
<td>97.92</td>
<td>2.84</td>
</tr>
<tr>
<td>SR (2) + TR (1.333)</td>
<td>82.29</td>
<td>93.75</td>
<td>6.77</td>
</tr>
<tr>
<td>SR (3) + TR (3.333)</td>
<td>68.75</td>
<td>86.46</td>
<td>12.88</td>
</tr>
<tr>
<td>SR (0) + TR (3.333)</td>
<td>93.75</td>
<td>97.91</td>
<td>2.32</td>
</tr>
<tr>
<td>SR (3) + TR (0)</td>
<td>84.38</td>
<td>95.83</td>
<td>7.58</td>
</tr>
</tbody>
</table>

Table 4. Identification and verification results obtained by fusioning with the adaptive weighting operator.

<table>
<thead>
<tr>
<th>Method</th>
<th>1-Best (%)</th>
<th>3-Best (%)</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR (0) + TR (0)</td>
<td>96.88</td>
<td>97.92</td>
<td>2.75</td>
</tr>
<tr>
<td>SR (1) + TR (0.333)</td>
<td>92.71</td>
<td>97.92</td>
<td>4.07</td>
</tr>
<tr>
<td>SR (2) + TR (1.333)</td>
<td>84.38</td>
<td>92.71</td>
<td>6.06</td>
</tr>
<tr>
<td>SR (3) + TR (3.333)</td>
<td>68.75</td>
<td>85.42</td>
<td>13.16</td>
</tr>
<tr>
<td>SR (0) + TR (3.333)</td>
<td>95.83</td>
<td>97.92</td>
<td>2.89</td>
</tr>
<tr>
<td>SR (3) + TR (0)</td>
<td>84.38</td>
<td>95.83</td>
<td>9.52</td>
</tr>
</tbody>
</table>

By looking at the second, third and fourth rows of each table, it is possible to appreciate the best results of the four configurations, because there is no artificial noise corrupting them (the noise value in parenthesis is 0). From the results it is clear that the multimodal integration of the spatial and temporal systems increases the correct identification scores; we also observed a similar behaviour for the verification task.

Finally, the last two rows of Table 3 and 4 represent the recognition scores for unbalanced multimodal systems, which are obtained by fusioning a performing one with a weak one. In the tables we report the results for extreme cases, those with the highest unbalance, in order to evaluate the degradation in the worst cases. From this experiments we can see that, even if the best multimodal systems are obtained by fusioning the best unimodal ones, the unbalanced fusion is not excessively degradating the scores of the best system; thus, our results support the argument that it is always better to fuse complementary systems, even if the have different performances.

4. CONCLUSION AND FUTURE WORKS

In this article, we analysed the effects of using at the same time the physiological and behavioural information, which is embedded in video data, for recognition purposes. Our experimental results show that this integration provides an important advantage in discriminating identities, and that in practical cases (those with balanced fusion) it is advantageous to fuse unimodal systems, especially in presence of corrupting noise. However, we are aware that our work needs a bigger experimental validation and that should be extended to diverse video databases.

Our system can be improved in multiple ways. One way could be to modify the Static Recognizer module, by replacing the PCA-based approach with a more performing recognition algorithm. Then, the Temporal Recognizer module might be improved by adding facial mimic information: it could integrate the eye blinking or the lip motion with the head displacements. Finally, there is a variety of fusion techniques which can be investigated and possibly applied to our multimodal approach.

5. REFERENCES