Advances on Feature Selection Techniques with Applications to Face Recognition

Teófilo E. de Campos[1]*, **Roberto M. Cesar Jr.**[2]

[1] Robotics Research Group – Department of Engineering Science
University of Oxford, Parks Road, Oxford, OX1 3PJ, United Kingdom

[2]Departamento de Ciência da Computação – Instituto de Matemática e Estatística
Universidade de São Paulo, Rua do Matão, 1010, 05508-090 São Paulo - SP - Brazil

teo@robots.ox.ac.uk, cesar@ime.usp.br
http://www.vision.ime.usp.br/~creativision

Abstract. *Although many efficient and robust algorithms have been developed for face recognition, there are still many challenges to be overcome. In particular, a robust and compact face representation is still to be found, which would allow fast recognition of different individuals. In order to address this problem, we assessed dimensionality reduction methods and introduced a new feature selection approach, which associates an efficient searching algorithm (sequential floating search methods) with a tolerance-based fuzzy distance. We verified the efficiency of this technique through exhaustive performance assessment experiments with synthetic and real data sets, the latter being obtained from a standard face recognition image database.*

1. Introduction

Person recognition from video sequences is an instigating research field that has attracted intense and growing attention from the vision research community, finding many important practical applications, such as in human-machine interaction, video indexing, and surveillance [Chellappa et al., 1995, Gong et al., 2000]. Most of such applications can only be suitably fulfilled by using a real-time, reliable and flexible implementation of a face recognition system. A project for the development of such a system regarding faces in video sequences has been developed at IME - USP exploring the framework showed in figure 1. The system architecture is divided in two parts, i.e. (1) face detection and tracking; and (2) data normalisation, feature extraction, and statistical pattern recognition. The former has been developed and described by another student. This paper describes a Master's thesis that has addressed some important problems regarding part (2), mainly with respect to feature selection.

There are some key problems to be solved in face recognition, such as the large volume of data, mainly for video sequence applications. In this context, an important issue

*The authors are grateful to FAPESP, CNPq, CAPES and CAPES-COFECUB projects for funding support, and to Isabelle Bloch for important discussions on fuzzy distances and feature selection.

for feature extraction and recognition is dimensionality reduction, because large dimensionality data can both dramatically increase the computational cost of the recognition task and decrease the correct recognition ratios by limiting the generalisation capabilities of the classifier [Jain et al., 2000]. Therefore, our thesis concentrated in the important issue of dimensionality reduction, which has several applications in pattern recognition problems, such as in data-mining, bio-informatics and multimedia databases, to name but a few. We have explored two strategies for dimensionality reduction: modular recognition followed by feature extraction and feature selection methods. Modular recognition consists of using image subregions, such as the eyes. In [de Campos et al., 2000a], we have shown that, when the training set is small, the correct classification rate is increased using eyes-only images. In some cases, the classification results obtained by using eye-only images were about 19% better than those using whole face images.

As far as feature selection methods are concerned, most of them are based on the optimisation of some criterion function using a search method. In this context, we have introduced a new criterion function, based on fuzzy clusters, which is more suitable for non-convex clusters, thus circumventing some problems with the standard approaches such as Mahalanobis distance. Our approach has been assessed through exhaustive experiments using simulated and real data and producing encouraging successful results, as described in the subsequent sections.

This paper is divided as follows. Section 2. presents a brief overview to the dimensionality reduction problem as a background to introduce our approach in Section 3.. Finally, Section 4. concludes this paper by discussing the experimental findings, contributions and future work to be developed from the work started in our thesis.

2. Dimensionality Reduction: Overview

Given a set X of unknown patterns and a set of known classes Ω, a classifier is a function $\Upsilon : X \rightarrow \Omega$, such that, given a pattern $\mathbf{x}_i \in X$, $\Upsilon(\mathbf{x}_i) = \omega_j$, in which ω_j is the j-th class of Ω. In the case of face recognition, Ω represents the known people whereas X contains the patterns $\mathbf{x}$ obtained from unknown face images. In this work, we have done experiments using both the minimum distance to the prototype and the k-nearest neighbour classifiers. A pattern is represented as a feature vector $\mathbf{x} = [x_1, x_2, \cdots, x_N]^t$ of dimensionality N. $\mathbf{x}$ is modelled as a random vector such that each *feature* x_j ($1 \leq j \leq N$) is a taken as a random variable [Duda and Hart, 1973]. There is no general rule about how to select the best features for a given pattern recognition task, which emphasises the importance of the role played by feature selection methods in order to avoid the so called *curse of dimensionality* [Jain et al., 2000].

Given a feature set $\mathbf{x}$, of dimensionality N, a feature extraction method $\mathcal{H}$ is a transformation $\mathcal{H} : \mathbf{x} \rightarrow \mathbf{y}$, such that the dimensionality of $\mathbf{y}$ is m, and usually $m < N$. Thus, given a pattern $\mathbf{x}_i$, $\mathcal{H}(\mathbf{x}_i) = \mathbf{y}_i$, and $\mathbf{y}_i$ is the new representation of $\mathbf{x}_i$ in the feature space defined by $\mathbf{y}$. As a transformation, feature extraction methods modify the feature space to represent the patterns. We have explored two different approaches for feature extraction, namely the Fourier transform and Principal Components Analysis (PCA).

On the other hand, automatic feature selection methods are important in applications where it is not desirable to modify the feature space, or when a feature extraction

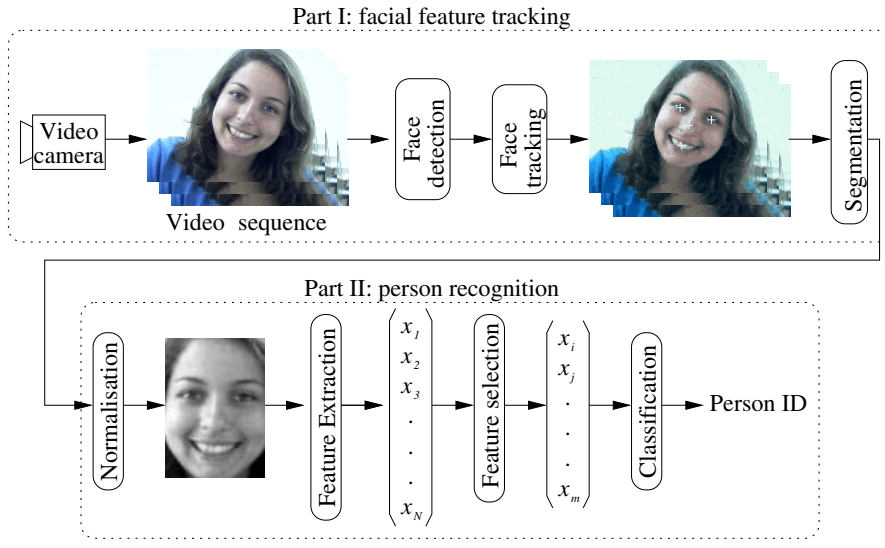

Figure 1: Framework for face recognition from video sequences. Part I: facial features location; part II: dimensionality reduction and recognition.

method has already been applied. Given a set $\mathbf{x}$ of N features, a feature selection algorithm tries to select a subset $\mathbf{y}$ such that $|\mathbf{x}| = m$, and $J(\mathbf{x}) = \max_{\mathbf{z}\subseteq\mathbf{y},|\mathbf{z}|=m} J(\mathbf{z})$, where $J(\cdot)$ is a criterion function. An example of $J(\cdot)$ that is often used for pattern recognition problems is the correct classification rate. Nevertheless, there are some situations in which this is not feasible in practise (e.g. when training a classifier is too time consuming, and when it is better not having a feature set fitted to a single specific classifier) and alternative criterion functions must be explored to determine the fitness of the feature set without the need for training and testing classifiers.

Several search methods for feature selection have been proposed. The optimal methods can be applied for small feature sets or when a monotonic criterion function is employed, i. e., when it can be ensured that there is no curse of dimensionality, which is rarely true in practise. According to [Jain et al., 2000], the sequential floating search methods (SFSM) and their adaptive versions (ASFSM) [Somol et al., 1999] are the best sub-optimal methods. For this reason, we have assessed these algorithms for face detection using Fourier descriptors [de Campos et al., 2000b], showing their superior performance with respect to other non-automatic approaches (the difference in correct classification rate is about 20%). The results obtained by the ASFSM are slightly better than those provided by the SFSM (about 4% better in the most discrepant case), but the time complexity of ASFSM is much higher than that of SFSM. In the same experiment, SFSM took about 2 seconds to run, while ASFSM took about 4 hours. This motivated our choice for the SFSM for the recognition experiments in the thesis.

3. A new criterion function

Most of the separability measures that are used as criterion functions are suitable for convex clusters and may fail on selecting feature sets that can provide good classification results for classifiers with more complex decision boundaries (i.e. non-linear), like k-nearest neighbours. In order to circumvent this problem, we have

evaluated the feature selection performance of several well-known distances between clusters. The survey in [Bloch, 1999] indicates that the fuzzy distance proposed in [Lowen and Peeters, 1997] is the most suitable separability measure for our application. Therefore, in [de Campos et al., 2001], we proposed a new criterion function based on that fuzzy measure, aiming at maximising the distance between the centroids of the classes and the compactness of each class, while minimising the amount of overlapped regions of the two classes in the feature space.

In order to calculate the criterion function, it is necessary to fuzzyfy the training samples of each class. We chose a fuzzification process that define the typicality of a pattern as being inversely proportional to its distance to the prototype of its class (i.e. the support of the cluster). The prototype was defined as the mean of the class, leading to:

$$\nu_\omega(\mathbf{x}_i) = \begin{cases} \frac{1}{1+d(\mathbf{x}_i, p_j^\omega)}, \mathbf{x}_i \in \omega, \\ 0, \mathbf{x}_i \notin \omega, \end{cases} \tag{1}$$

for $j = 1, 2, \ldots, \mathcal{P}$, where $\mathbf{x}_i$ is a pattern, $\nu_\omega(\mathbf{x}_i)$ is the membership function of that pattern to the set ω, p_j^ω represents the jth support of the cluster ω, $d(\cdot)$ is the Euclidean distance, and $\mathcal{P}$ is the number of possible supports per cluster. In our experiments, we used only one support per cluster ($\mathcal{P} = 1$). The separability measure that we used is defined upon a local (regarding to a pattern $\mathbf{x}$) difference between two classes defined as:

$$d_{\mathbf{x}_i}^\tau(\nu_{\omega_r}, \nu_{\omega_s}) = \inf_{\mathbf{x}_j, \mathbf{x}_k \in B(\mathbf{x_i}, \tau)} |\nu_{\omega_r}(\mathbf{x}_j) - \nu_{\omega_s}(\mathbf{x}_k)|, \tag{2}$$

where $B(\mathbf{x}, \tau)$ denotes a N-dimensional ball having radius τ, centred in $\mathbf{x}$, being responsible for measuring the amount of overlapping regions between the classes. The τ parameter is called tolerance of the measure and $\nu_{\omega_q}(\mathbf{x}_i)$ is the membership function of the pattern $\mathbf{x}_i$ to the class q. Therefore, the measure is defined as

$$d_p^\tau(\nu_{\omega_r}, \nu_{\omega_s}) = [\int_{\mathcal{F}} [d_{\mathbf{x}}^\tau(\nu_{\omega_r}, \nu_{\omega_s})]^p d\mathbf{x}]^{1/p}, \tag{3}$$

where $\mathcal{F}$ represents the whole feature space. As this criterion function is based on a fuzzy measure, it can be adapted to any sort of data by changing the fuzzification function. This would have direct influence on how the aspects of the clusters influence the criterion function results. We have successfully assessed the introduced function using a database with 20.000 simulated patterns, with the complete results being described in [de Campos et al., 2001].

4. Concluding Remarks

We have developed, implemented and assessed several dimensionality reduction strategies. Our start point was the use of modular images for recognition, which was followed by the assessment of feature extraction and feature selection methods, including PCA, Fourier descriptors, feature selection and search methods, and criterion functions. In [de Campos et al., 2000a], we have shown that eyes-only images can provide better classification results than whole face images. The SFSM and the ASFSM feature selection algorithms have been assessed for face detection using Fourier descriptors in

[de Campos et al., 2000b]. This work has shown the superior performance of such feature selection algorithms with respect to other non-automatic approaches.

The SFSM algorithm has been explored together with the fuzzy measure explained in Section 3. providing successful results. This work showed that our approach circumvents a drawback of the traditional cluster distances, which are more suitable for convex clusters. Finally, we expanded the criterion function discussed above to create a version that works for problems having more than two classes. This is an essential issue, since our aim was to apply it for face recognition problems, in which usually there are much more than two classes. The new criterion function has been used to perform exhaustive feature selection experiments varying a tolerance parameter τ. The feature subsets obtained by our method over performed most of the classifier-based feature selectors that we assessed in a normalised eigeneyes space for person recognition. Our results are promising and feature selection method is currently being further developed by other researchers at IME-USP in order to be applied in bio-informatics and shape analysis problems.

References

Bloch, I. (1999). On fuzzy distances and their use in image processing under imprecision. *Pattern Recognition*, 11(32):1873–1895.

Chellappa, R., Wilson, C. L., and Sirohey, S. (1995). Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):703–740.

de Campos, T. E., Bloch, I., and Cesar-Jr, R. M. (2001). Feature selection based on fuzzy distances between clusters: First results on simulated data. In *Lecture Notes in Computer Science*, volume 1973, pages 186–195, Rio de Janeiro, Brasil. Springer-Verlag Press.

de Campos, T. E., Feris, R. S., and Cesar-Jr., R. M. (2000a). Eigenfaces versus eigeneyes: First steps towards performance assessment of representations for face recognition. In *Lecture Notes in Artificial Intelligence*, volume 1973, pages 197–206, Acapulco, Mexico. Springer-Verlag Press.

de Campos, T. E., Feris, R. S., and Cesar-Jr, R. M. (2000b). Improved face $\times$ non-face discrimination using fourier descriptors through feature selection. In *13th SIBGRAPI*, pages 28–35. IEEE Computer Society Press.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley-Interscience, USA, 1st edition.

Gong, S., McKenna, S., and Psarrou, A. (2000). *Dynamic Vision: From Images to Face Recognition*. Imperial College Press, UK.

Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37.

Lowen, R. and Peeters, W. (1997). On various classes of semi-pseudometrics used in pattern recognition. In *7th IFSA World Congress*, volume I, pages 232–237, Prague.

Somol, P., Pudil, P., Novovicová, J., and Paclík, P. (1999). Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20:1157–1163.