Classification-Specific Feature Sampling for Face Recognition

Effrosyni Kokiopoulou and Pascal Frossard
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Institute - ITS
CH - 1015 Lausanne, Switzerland
Email: {effrosyni.kokiopoulou,pascal.frossard}@epfl.ch

***Abstract*— Feature extraction based on different types of signal filters has received a lot of attention in the context of face recognition. It generally results into extremely high dimensional feature vectors, and sampling of the coefficients is required to reduce their dimensionality. Unfortunately, uniform sampling that is commonly used to that aim, does not consider the specificities of the recognition task in selecting the most relevant features. In this paper, we propose to formulate the sampling problem as a supervised feature selection problem where features are carefully selected according to a well defined discrimination criterion. The sampling process becomes specific to the classification task, and further facilitates the face recognition operations. We propose to build features on random filters, and Gabor wavelets, since they present interesting characteristics in terms of discrimination, due to their high frequency components. Experimental results show that the proposed feature selection method outperforms uniform sampling, and that random filters are very competitive with the common Gabor wavelet filters for face recognition tasks.**

I. Introduction

Face recognition is one of the most challenging tasks in computer vision and image processing with numerous applications including security biometric systems, surveillance and human machine interfaces, only to name a few. The main difficulties arise from the different appearance that a face may have under different illumination conditions, facial expressions and various poses. The reader is referred to [1] for more information about the face recognition challenges and its applications.

Most of the recognition algorithms build sets of features that correspond to the response of well-chosen filters positioned over the face images. It may result in a very large vector of coefficients, which is usually uniformly (down)sampled in order to reduce its dimensionality. However, such a sampling process is oblivious of the subsequent face recognition task. In order to alleviate this problem, the authors in [2] propose an adaptive sampling scheme which is based on intuitive arguments that the high variance coefficients are the most important ones for the recognition task. They use a threshold on variance which is obtained from training data. However, their scheme does not exploit the available class labels of the training data.

In this paper, we propose to formulate the sampling process as a supervised feature selection problem [8, ch.9]. We provide a suboptimal greedy algorithm that finds a set of features according to a well defined class separability cost function. In each step the algorithm greedily selects the feature that results in the highest gain in terms of discrimination. The cost function that we use is the ratio of inter-class variance over the intra-class variance. We also show that the evaluation of the cost function can be done efficiently, and we propose a fast feature selection algorithm that is adapted to the face recognition task.

Feature extraction is based on the popular Gabor wavelets (filters), that have been used in many face recognition systems (e.g., [2], [4], [5], [6] and references therein). Gabor wavelets are known to have nice spatial frequency characteristics and orientation selectivity. As an alternative, we propose the use of random filters for feature extraction. The entries of such filters are i.i.d. random variables drawn independently from the Bernoulli/Rademacher distribution of $\{\pm 1\}$'s. Random filters have been successfully applied for compressed sensing and signal reconstruction [3]. Interestingly, the quality of the features extracted from the random filters are shown to be competitive to that of Gabor filters. This suggests that in general the oscillatory nature or high frequency response of a filter is one of its most crucial characteristics in terms of discrimination. Finally, experimental results demonstrate that the classification-specific feature selection algorithm clearly outperforms the uniform sampling of filters coefficients.

II. Recognition-specific sampling of feature coefficients

Recognitions systems commonly represent face images as a large set of features, which generally represent the responses of well-chosen filters at predefined pixel positions. However, all features do not contribute equally to the face recognition task. Therefore, in order to reduce the dimensionality of the image feature vector, we propose to formulate the sampling process as a feature selection problem [8, ch.9], illustrated in Figure 1. The features sampling process thus becomes supervised, and more effective for discrimination purposes.

The feature selection seeks for the optimal set of d features out of m. One possible approach would be to do an exhaustive search among all $\begin{pmatrix} m \\ d \end{pmatrix}$ possible feature sets and choose the best one (according to the discrimination criterion at hand). However, such an approach is computationally very expensive. The reader is referred to [8, ch.9] for more information

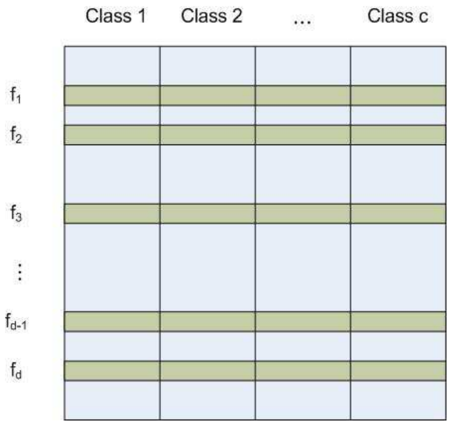

Fig. 1. The feature selection process. Columns correspond to data samples and rows correspond to features.

about the various feature selection methods. In this paper, we propose a suboptimal greedy algorithm which retains an active set of features. In each step, the algorithm greedily selects the feature that results in the highest increase in the separability cost function and adds it to the active set of features. The cost function that we use is the inter-class variance over the intra-class variance.

In particular, consider the data matrix $X \in R^{m\times n}$. We denote the number of classes by c and assume without loss of generality that $X = [X^{(1)}, \dots, X^{(c)}]$, where $X^{(i)} \in R^{m\times n_i}$ denotes the data samples that belong to the i-th class of cardinality n_i. Note that in the context of face recognition, each class corresponds to a different individual. We also denote $\mu^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_j^{(i)}$ the centroid of the i-th class and $\mu = \frac{1}{n}\sum_{j=1}^{n} x_j$ the global centroid. The within-class scatter matrix S_w and between-class scatter matrix S_b are defined as follows,

$$S_w = \frac{1}{n}\sum_{i=1}^{c}\sum_{x\in X_i}(x-\mu^{(i)})(x-\mu^{(i)})^\top \quad (1)$$

$$S_b = \frac{1}{n}\sum_{i=1}^{c} n_i(\mu^{(i)}-\mu)(\mu^{(i)}-\mu)^\top. \quad (2)$$

The cost function that we use for feature selection is defined as follows

$$J = \frac{\mathrm{tr}(S_b)}{\mathrm{tr}(S_w)}. \quad (3)$$

In order to obtain a computationally efficient algorithm, it is crucial to have a fast evaluation of the cost function, for each candidate feature set. We therefore propose an efficient way of evaluating the cost function J. First, denote $e^{(i)} = [1, \dots, 1]^\top \in R^{n_i\times 1}$ and $e = [1, \dots, 1]^\top \in R^{n\times 1}$. If we further define the matrices

$$\begin{aligned} G_w &= \frac{1}{\sqrt{n}}[X^{(1)} - \mu^{(1)}(e^{(1)})^\top, \dots, X^{(c)} - \mu^{(c)}(e^{(c)})^\top] \\ G_b &= \frac{1}{\sqrt{n}}[\sqrt{n_1}(\mu^{(1)} - \mu e^\top), \dots, \sqrt{n_c}(\mu^{(c)} - \mu e^\top)], \end{aligned}$$

then we observe that $S_w = G_w G_w^\top$ and $S_b = G_b G_b^\top$. Interestingly, we notice that

$$\begin{aligned} \mathrm{tr}(S_w) &= \mathrm{tr}(G_w G_w^\top) = \mathrm{tr}(G_w^\top G_w) = \frac{1}{n}\sum_{i=1}^{c}\sum_{x\in X_i}\|x-\mu^{(i)}\|_2^2, \\ \mathrm{tr}(S_b) &= \mathrm{tr}(G_b G_b^\top) = \mathrm{tr}(G_b^\top G_b) = \frac{1}{n}\sum_{i=1}^{c} n_i\|\mu^{(i)}-\mu\|_2^2. \end{aligned}$$

The above property allows to build a very efficient algorithm for the cost function evaluation. In particular, we make use of the following lemmas to simplify the calculation of the cost function J for each new candidate feature set.

Lemma 1: If the data are partitioned row-wise in two parts i.e.,

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \mu^{(i)} = \begin{bmatrix} \mu_1^{(i)} \\ \mu_2^{(i)} \end{bmatrix},$$

then it holds that $\mathrm{tr}(S_w) = \mathrm{tr}(S_w^{(1)}) + \mathrm{tr}(S_w^{(2)})$, where $S_w^{(1)}$ and $S_w^{(2)}$ are the within-class scatter matrices induced by the partition.

Proof: Notice that

$$\begin{aligned} \mathrm{tr}(S_w) &= \frac{1}{n}\sum_{i=1}^{c}\sum_{x\in X_i}(x-\mu^{(i)})^\top(x-\mu^{(i)}) \\ &= \frac{1}{n}\sum_{i=1}^{c}\sum_{x\in X_i}\{(x_1-\mu_1^{(i)})^\top(x_1-\mu_1^{(i)}) + \\ &\quad (x_2-\mu_2^{(i)})^\top(x_2-\mu_2^{(i)})\} \\ &= \frac{1}{n}\sum_{i=1}^{c}\sum_{x\in X_i}(x_1-\mu_1^{(i)})^\top(x_1-\mu_1^{(i)}) + \\ &\quad \frac{1}{n}\sum_{i=1}^{c}\sum_{x\in X_i}(x_2-\mu_2^{(i)})^\top(x_2-\mu_2^{(i)}) \\ &= \mathrm{tr}(S_w^{(1)}) + \mathrm{tr}(S_w^{(2)}). \end{aligned}$$

■

We also provide a second lemma for the case of S_b.

Lemma 2: According to the assumptions of Lemma 1, it holds that $\mathrm{tr}(S_b) = \mathrm{tr}(S_b^{(1)}) + \mathrm{tr}(S_b^{(2)})$, where $S_b^{(1)}$ and $S_b^{(2)}$ are the between-class scatter matrices induced by the partition.

Proof: The proof is similar to the proof of Lemma 1 and is omitted. ■

The proposed bottom-up feature selection algorithm is finally summarized in Table I. Initially, it starts with an empty active set S. In each step, the feature which results in the highest increase in the value of J is selected and added to the active set.

III. Features for face recognition

A. Gabor wavelets

The 2D Gabor wavelets (or filters) are commonly used in face recognition algorithms, due to their good discrimination properties. They are defined as follows

$$\phi_j(\vec{x}) = \frac{\|\vec{k}_j\|_2^2}{\sigma^2}\exp(-\frac{\|\vec{k}_j\|_2^2\|\vec{x}\|_2^2}{2\sigma^2})[\exp(i\vec{k}_j\vec{x}) - \exp(-\frac{\sigma^2}{2})], \quad (4)$$

```
Algorithm A1: Feature Selection
Input: $X \in R^{m \times n}$: Data matrix,
$d$: number of features selected.
Output: $S$: set of selected features.
1. Compute the value of $J$ for each feature individually.
2. $S = \emptyset$ (active set), $k = |S| = 0$ (number of selected features).
3. while $k < d$
4.     For each feature $i \notin S$ , compute the value of
       $J(S'),\ S' = S \bigcup i$.
5.     Find the feature $j$ resulting in the maximum value of $J$.
6.     Include this feature in the active set, $S = S \bigcup \{j\}$ .
7.     $k = k + 1$.
4. end.
```

TABLE I

THE FEATURE SELECTION ALGORITHM.

where $\vec{k}_j = k_p e^{i\theta_q}$, with $k_p = \frac{0.5\pi}{(\sqrt{2})^p}$ and $\theta_q = q\frac{\pi}{8}$. The Gabor wavelet basically consists of a Gaussian envelope modulated by a complex exponential, given in the first term in the brackets of equation (4). The second term in the brackets is included in order to make the wavelet zero mean. In common applications that work with small face images, Gabor wavelets are sampled at five scales $p = 0, \ldots, 4$ and eight orientations $q = 0, \ldots, 7$. This results in a filter bank of $p \times q = 40$ filters (wavelets), which is depicted in Figure 2.

Feature extraction is done by convolving the image $I(\vec{x})$ with each one of the wavelets in the filter bank.

$$F_j(\vec{x}) = I(\vec{x}) * \phi_j(\vec{x}), \tag{5}$$

where $*$ denotes the convolution operator and $\vec{x} = (x, y)$. This process results into a vector field (one feature vector of length 40 for each pixel) which is then reshaped to yield a high dimensional vector of length $40 \cdot n_x \cdot n_y$, where n_x and n_y represent the size of the image.

B. Random filters

As an alternative to the Gabor filters, we propose to make use of random filters, which typically present high frequency characteristics and have been successfully applied to compressed sensing and signal reconstruction [3]. The entries of the random filters are i.i.d. random variables drawn independently from the Bernoulli/Rademacher distribution of $\{\pm 1\}$'s (see Figure 2). Although one may expect that the convolution of an image with a random filter will result into some kind of random patterns, it surprisingly turns out that this process preserves enough information of the image, which is useful for discrimination. We show that the random filters provide competitive results with those offered by Gabor wavelets. This suggests that the most crucial characteristic of Gabor filters in terms of discrimination, is their oscillating or high frequency part.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

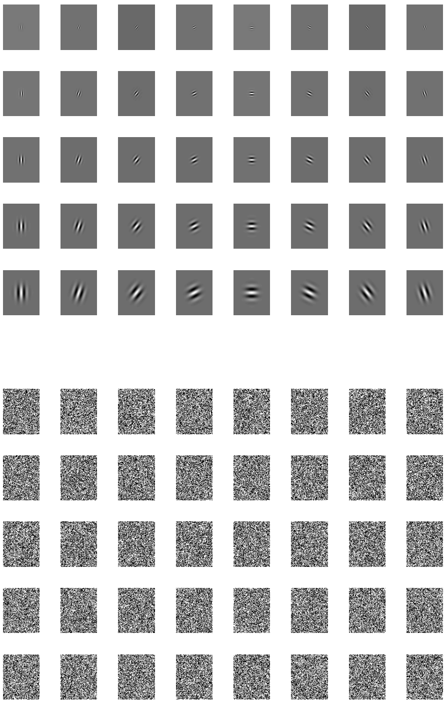

Fig. 2. Bank of (top) Gabor wavelets filters (real part) and (bottom) random filters.

Fig. 3. Sample face images from the YALE database.

For our experimental evaluations, we use the YALE database [7] which consists of 15 subjects including 11 images per subject (165 images in total). The database contains variations in lighting as well as in facial expressions. Figure 3 depicts a few sample face images from the YALE database. In the preprocessing stage, the images are closely cropped so that they contain only the main part of the face, and then they are resized to dimension 32×32.

We report the classification error rate by measuring the leave-one-out (LOO) error which is computed as follows: a facial image is used as a probe image and then classified by nearest neighbor (NN) rule, using as training data the remaining images in the collection. This is repeated for every facial image in the collection. The LOO error rate is finally the percentage of misclassified facial images.

It can be noted that the NN rule used in the experiments is a simple classification rule, but not necessarily the best one. Note that the feature selection could be subsequently combined with a dimensionality reduction method such as Principal Component Analysis (PCA) [8, ch.9] or Linear Discriminant Analysis (LDA) [8, ch.4] that could potentially improve the recognition rate. However, this is out of the scope of the present paper, where our focus is on the sampling/feature selection stage, demonstrated in the context of face recognition.

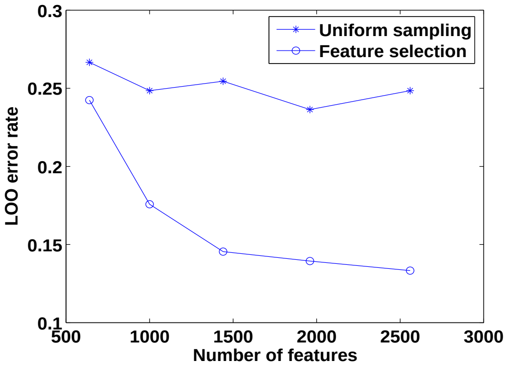

Fig. 4. LOO error rate (measured on Gabor features) of uniform sampling vis-a-vis feature selection.

a) Feature selection vs uniform sampling: In the first experiment we use the Gabor features and compare the proposed feature selection method with the uniform sampling strategy. The initial dimensionality of the feature vector is $32 \cdot 32 \cdot 40 = 40960$. The uniform sampling is implemented by sampling the pixel feature vectors with the uniform pattern $[1 : k : 32, 1 : k : 32]$ (in MATLAB notation) for $k = \{4, 5, 6, 7, 8\}$. This results into 2560, 1960, 1440, 1000 and 640 total number of features respectively. We compare the uniform sampling with the proposed feature selection algorithm for the same total number of features, and the LOO error rates are shown in Figure 4. It is clear that the supervised feature selection process is superior to the naive uniform sampling. This is reasonable since the former is supervised, and optimized to increase the classification performance of the selected features.

b) Random filters: In the second experiment we study the performance of the random filters in terms of LOO error rate. The features in this case are the convolution coefficients of the image with a filter bank of 40 random filters. We measure the error rate of both uniform sampling and feature selection on the extracted features. Note that the random filters have many possible realizations. Thus, in order to remove any bias in our measurements, we repeat the experiment 10 times and measure the LOO error rates for different random realizations of the filter bank. We report the results in boxplot notation in Figure 5. The boxes have lines at the lower quartile, median, and upper quartile values. Observe again that feature selection is superior to uniform sampling, as it was the case with the previous experiments using the Gabor features. Observe also that the features extracted from the random filters are competitive to those extracted from Gabor wavelets. Interestingly, for small values of feature set size, they seem to behave even better. This experiment suggests that filters with high frequency characteristics yield features that facilitate the recognition task.

V. Conclusions

This paper has proposed a new method for the sampling of high dimensional feature vectors for face recognition. The sampling process is formulated as a supervised feature selection problem where the features are selected according to their discriminating value. Additionally, we have described a greedy algorithm with an efficient implementation which is based on fast evaluations of the class separability cost function. Such a selection process has been shown to outperform common uniform sampling strategies. Finally, we propose to build features on random filters, whose high frequency characteristics offer performance competitive to those of common Gabor filters.

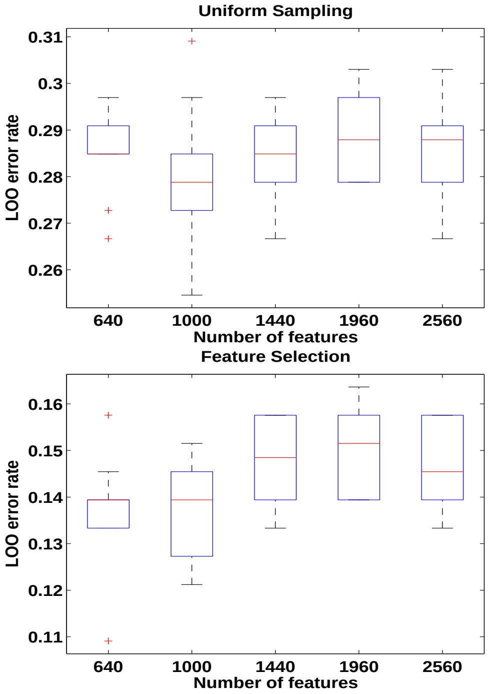

Fig. 5. LOO error rate of features extracted from random filters. Top panel: uniform sampling and bottom panel: feature selection.

References

[1] W. Zhao, R. Chellapa, P. Phillips, and A. Rosenfeld, "Face Recognition: a Literature Survey" ACM Computing Surveys, vol. 35(4), pp 399458, December 2003.

[2] S. Du and R. Ward, "Statistical Non-uniform Sampling of Gabor Wavelet Coefficients for Face Recognition", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), March. 2005, Philadelphia, PA, USA.

[3] J. Tropp, M. Wakin, M. Duarte, D. Baron and R. Baraniuk, "Random Filters for Compressive Sampling and Reconstruction", to appear in IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), May. 2006, Toulouse, France.

[4] L. Wiskott, J. M. Fellous and C. Von Der Malsburg, "Face Recognition by Elastic Bunch Graph Matching", IEEE Trans. Patt. Anal. Mach. Intell., vol. 19, pp. 775-779, 1997.

[5] C. Liu and H. Wechsler, "Independent Component Analysis of Gabor Features for Face Recognition", IEEE Trans. Neur. Net., vol. 14(4), pp. 919-928, 2003.

[6] C. Liu, "Gabor-based Kernel PCA with Fractional Polynomial Models for Face Recognition", IEEE Trans. Patt. Anal. Mach. Intell., vol. 26(5), pp. 572-581, 2004.

[7] P. Belhumeur, J. Hespanha and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", IEEE Trans. Patt. Anal. Mach. Intell., Special Issue on Face Recognition , vol. 19(7), pp. 711-720, July 1997.

[8] A. Webb, "Statistical Pattern Recognition", Wiley, 2002, 2nd edition.