Abstract

Biologically inspired model (BIM) is proven to be an effective feature representation approach for visual object categorization. In BIM, two successive S(simple)-to-C(complex) hierarchical layers are performed to simulate the visual perception process of primate visual cortex. However, the intensive computational cost above C1 layer in BIM extremely limits its application in real-time object recognition tasks. This paper proposes to use a set of improved early biologically inspired features (EBIF, including S1 and C1) for face recognition, in which pyramidal statistics of mean and standard deviation rather than MAX pooling are used for scale-tolerant feature condensation and local normalization is performed on C1 layer. Incremental PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) are then combined to efficiently learn a discriminant subspace for feature dimensionality reduction. In the matching stage, Cosine similarity is adopted as the distance metric for a given face pair. Experimental results on two public face datasets and a mobile face dataset show the effectiveness of the proposed method.

1. Introduction

Face recognition [6, 26] has received significant attention during last decades due to its wide range of applications, such as customs clearance, human-computer interaction, visual surveillance, and so on. With the surging growth of smart mobile devices, mobile technology is changing people’s life in entertainment, shopping, social network and so on. It can be predicted that face recognition, as a convenient authentication method on mobile devices, will play a very important role in securing our money/assets and protecting our privacy in the world of mobile Internet.

Face recognition is still a challenging problem due to appearance variations caused by illumination, pose, facial expressions, and so on. How to represent a face in an appropriate way is the key to face recognition. Modern face recognition methods can be generally divided into two categories: holistic appearance based methods[12, 21, 1, 27, 25]. Holistic appearance based approaches usually construct a subspace using Principal Component Analysis (PCA) [19, 7], Linear Discriminant Analysis (LDA) [3, 5, 22], or Independent Component Analysis (ICA) [2]. The face images are then projected and compared in a low-dimensional subspace in order to avoid the curse of dimensionality. Holistic appearance based representation usually performs well in controlled environment with a large and representative training set.

Recent studies [27, 1, 9, 25] show that local appearance based approaches usually outperform holistic approaches because of their ability to deal with appearance variations locally. Pentland et al. [12] extended the eigenface [19] technique to a layered representation by combining eigenfaces and other eigenmodules, such as eigeneyes, eigen noses, and eigenmouths. Wiskott et al. [21] achieved good face recognition performance with the Elastic Bunch Graph Matching (EBGM) method which represents a face by a graph-based model. Local Binary Pattern (LBP), which was originally designed for texture classification [11], was introduced in face recognition in [1] and found performing much better than eigenface and EBGM. In recent years, Gabor wavelet [9, 27] features have been successfully used in face recognition because of their excellent performance on orientation and spatial frequency selectivity. Zhang et al. [25] proposed Local Gabor Binary Pattern Histogram Sequence (LGBPHS) for face recognition by combining Gabor filters and the local binary operator, resulting in better performance than LBP and Fisherface [3]. Chai et al. [4] proposed to exploit Gabor features together with Ordinal Measures [18] for face recognition, which shows better performance than LGBPHS.

This paper exploits Early Biologically Inspired Features (EBIF) for face recognition. Biologically Inspired Model (BIM) [14, 17, 10] is proven to be an effective representation approach for visual object categorization. The standard model of BIM consists of four layers of computational units where simple S units alternate with complex C units. BIM usually exhibits a good trade-off between invariance and selectivity because features in BIM are slightly
position- and scale-tolerant over neighboring positions and multiple orientations (like complex cells in primary visual cortex). However, the high layers (S2 and C2) of BIM require a large number of random prototype patches (how to obtain is tricky) used as visual vocabulary and need to compute millions of Gaussian-tuning distances between exhaustive C1 patches and the vocabulary, resulting in prohibitive computational cost for real-time tasks. Moreover, although the “bag-of-feature” style representation in the high layers is good for natural image categorization (has several tens of images per category for training), it may not be an appropriate way for face recognition (usually has several or even only one image per person). Therefore, we discard the above two high layers (S2 and C2) and improve the early biologically inspired features (S1 and C1) for face recognition by utilizing more robust statistical features and employing local normalization in C1 layer. To reduce irrelevant features in EBIF, incremental PCA (Principal Component Analysis) [8] and Linear Discriminant Analysis (LDA) are combined to learn a subspace efficiently. Experimental results show that the proposed EBIF descriptor outperforms some state-of-the-art face recognition methods on the FERET dataset [13]. We also validate the proposed method on a mobile face dataset for face authentication on mobile devices.

The remainder of this paper is organized as follows. Section 2 presents the overview of the proposed method. In Section 3, face representation using EBIF is described. Subspace learning for feature dimension reduction is presented in Section 4. Experimental results and analysis are presented in Section 5. Finally, we draw our conclusions in Section 6.

2. Overview of the proposed method

Figure 1 shows the flowchart of our EBIF based face recognition method. In the EBIF feature extraction process, an image pyramid with N scales is created for the input face image. For any two adjacent scales k and k+1 (k=1,2,...,N-1) in the image pyramid, a Gabor filter bank with M orientations is employed to extract low-level orientation information at each location of the pyramid, resulting in M Gabor pyramids in the S1 layer. In C1 layer, each Gabor pyramid is evenly divided into a set of sub-pyramids, and scale-fused pyramidal statistics are computed inside each sub-pyramid to measure the statistical distribution of object edge information along a specific orientation at a local position. To make feature set robust to illumination variations, we locally normalize those statistical features of all orientations inside a sub-pyramid by L2-norm. Then, those locally normalized multi-orientation scale-fused statistical features of all sub-pyramids are concatenated to represent the face image at scale k (we call them feature subsets at scale k). Finally, all the N-1 feature subsets constitute a multi-scale multi-orientation and slightly position- and scale-tolerant descriptor for face representation. EBIF descriptor is a rich but high-dimensional (the dimension is larger than 25,000) feature descriptor containing many irrelevant features that could harm both the accuracy and efficiency of the matching process, thus PCA-LDA is performed on the training set to learn a low-dimensional subspace for feature dimension reduction. Since directly computing the PCA subspaces from a large training set in a high-dimensional feature space is quite inefficient as well as expensive in memory-consuming, we introduce an incremental PCA [8] method to learn the PCA subspace efficiently. Then, LDA subspace can be computed efficiently from the projected features in the PCA subspace. In the matching process, we use the cosine value to measure the similarity of two feature templates.

3. EBIF

Humans and primates outperform the best machine vision systems, and building a vision system that can emulate object recognition in visual cortex has always been a dream of neuroscientists and computer vision technologists. The early work of Biologically Inspired Model (BIM) can be traced back to [14], which consists of four layers of computational units: S1, C1, S2 and C2, where S and C stand for simple and complex cells in the visual cortex. Simple S units combine their inputs with Gaussian-like tuning to increase object selectivity, while complex C units pool their inputs through a maximum operation, thereby introducing gradual invariance to scale and translation. The standard BIM model [14] has been able to quantitatively duplicate the generalization properties exhibited by neurons in inferotemporal monkey cortex. Serre et al. [17] tried to bridge the gap between computer vision and neuroscience, and extended the standard model to deal with real-world object recognition tasks (e.g., large scale natural image databases). Mutch and Lowe [10] further extended the base model to five layers including an additional initial image layer and improved the base model by increasing feature sparsity and retaining some spatial information above the intermediate feature level (S2).

In high layers of BIM model, usually several thousands of prototype patches (with various sizes: 4x4, 8x8 and etc) are randomly chosen over different scales of the C1 layers of random training samples and used as a visual dictionary. Each location of S2 layer is a dP-dimensional (dP is the number of prototype patches) Gaussian responses between the patches centered at this location and those random prototype patches. In the C2 layer, max pooling is performed for each prototype over all locations of S2 layer, resulting in a dp-dimensional “bag-of-feature” style feature descriptor. The extremely intensive computational cost (need to compute millions of patch-pair Gaussian-tuning
distances between C1 units and the dictionary) in S2 layer limits its application in real-time tasks. Moreover, the generation of random prototypes is very tricky (e.g., how to determine the numbers of prototypes over scales, object classes, locations, and etc) and “bag-of-feature” method needs tens of training images per class which usually cannot be satisfied in face recognition. Therefore, we discard the high layers (S2 and C2) of BIM, and improve the low layers of BIM for face representation.

Our early biologically inspired features (EBIF) are computed hierarchically in four layers: image layer, S1 layer, C1 layer, and local L2-norm layer. Figure 1 shows the flowchart of computation of the EBIF feature and details are given as follows.

**Image Layer**: The input face image is converted to grayscale and resized to $112 \times 112$ pixels. Then an image pyramid of $N$ (here $N=6$) scales is created, in which the adjacent higher scale is a factor of $2^{\frac{1}{2}}$ smaller than the current scale. The size of the 3D image pyramid will influence the recognition performance. Pyramid with larger bottom and more scales can capture more spatial and scale information from an object, but will result in more computational cost in the following layers. Our choice of the pyramid size is a tradeoff between the recognition accuracy and the computational cost. For each two adjacent scales $k$ and $k+1$ ($k=1,2,\ldots,N-1$) in the image pyramid, we obtain a feature subset to represent object at scale $k$ by performing the following steps.

**S1 Layer**: The units in the S1 layer correspond to simple cells in the visual cortex. In this layer, a Gabor filter bank is used to convolve the image layer at each position and scale of the image pyramid. Gabor filters are widely used in object recognition [9] because of their excellent performance on orientation and spatial frequency selectivity. A Gabor filter is the product of an elliptical Gaussian envelop and a harmonic function:

$$G(x, y) = \exp(-\frac{X^2 + \gamma^2 Y^2}{2\sigma^2})\cos\left(\frac{2\pi}{\lambda}X\right)$$  (1)

where $X = x \cos \theta + y \sin \theta$, $Y = -x \sin \theta + y \cos \theta$, and $\theta$ controls the orientation of the filter. Our filter bank consists of $M$ (here $M=9$) Gabor filters with $M$ orientations evenly distributed over $[0, \pi)$. Each filter is $5 \times 5$ in size and $x$ and $y$ vary between $-2$ and $2$. The parameters $\gamma$ (aspect ratio), $\sigma$ (effective width) and $\lambda$ (wavelength) are set to 1.0, 3.0, 5.0 respectively. The response of an image $I$ to a Gabor filter $G$ is given by:

$$R = |G * I|$$  (2)

The responses of the Gabor filter bank can localize object’s distinctive structures in different orientations, for example, in Figure 1, the S1 layer captures rich spatial frequency information around the eyes, nose, and mouth of the face in the image pyramid. Nonetheless, their spatial distribution is often not smooth. Direct use of features in the S1 layer for object recognition will suffer from appearance variations.

**C1 Layer**: This layer corresponds to complex cells in the visual cortex, which respond to oriented bars or edges anywhere within their receptive filed and are more broadly tuned to spatial frequency than simple cells. We evenly divide each Gabor pyramid (corresponding to a specific orientation) in S1 layer into a set of sub-pyramids with the size of $4 \times 4$ at the bottom and $3 \times 3$ on the top. In the base model of BIM, max pooling operation is performed in every sub-pyramid. The response of the max pooling is the maximum value of the units that fall in that sub-pyramid. As we know, max filtering is sensitive to image noise and illumination.
dimensions, so it is not a good statistical feature for visual recognition task with pool signal quality. Therefore, in our EBIF feature model, we utilize the combination of pyramidal statistics of mean and standard deviation rather than maximum response to represent each sub-pyramid. Experimental evidences will show that the new pyramidal statistical filtering method outperforms the max pooling operation significantly on face recognition.

The major advantage of C1 layer is that each element in this layer is a feature obtained by combining the response of local Gabor filter based edge-detectors that are slightly position- and scale-tolerant over neighboring positions and different orientations.

**Local L2-norm Layer:** Illumination-invariance is a very important property for face recognition tasks, especially for face recognition in mobile environment. To make EBIF robust to illumination variations, we normalize the $2 \times M$ (two statistics for each of the $M$ orientations) dimensional feature vector $f$ at each location of C1 layer by L2-norm:

$$\tilde{f} = \frac{f}{\sqrt{||f||^2 + \xi}}$$  \hspace{1cm} (3)

Where $\xi$ is a constant (here it is set to 0.001). We concatenate the normalized feature vectors at different locations of C1 layer to form a feature subset for object representation at scale $k$. All the $N-1$ multi-scale feature subsets, which contain local normalized multi-orientation edge information and are slightly position- and scale-tolerant, form the final EBIF feature representation.

## 4. Subspace learning

EBIF feature representation is in a high-dimensional feature space (more than 25,000), which may contain many irrelevant features that can harm both the accuracy as well as efficiency in the matching process. Thus, PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) are combined to learn a discriminative subspace for feature dimension reduction. However, traditional PCA computes the subspace in a batch mode, which is not efficient as well as problematic when the size of training data is very large (e.g., encounter out-of-memory problem). Thus, we introduce an incremental PCA method to efficiently learn the PCA subspace.

IPCA (incremental principal component analysis), known as SKL (Sequential Karhunen-Loeve) [8], is commonly used to incrementally learn a low dimensional eigenspace over a high-dimensional feature space.

**SKL algorithm** Assume previous observation data is a $d \times n$ matrix $O = \{o_1, o_2, \ldots, o_n\}$ (where $o_i$ is a $d$-dimensional feature vector) and $B$ is a $d \times m$ matrix of new observations, the SKL algorithm aims to compute the SVD (singular value decomposition): $[O \ B] = U\Sigma V'^T$ efficiently from the SVD: $O = U\Sigma V'^T$. Denote $\tilde{B}$ as the component of $B$ orthogonal to $U$: $\tilde{B} = B - U{\tilde{U}}^T B$, then:

$$[ O \ B ] = \begin{bmatrix} U & \tilde{B} \\ \Sigma \end{bmatrix} \begin{bmatrix} U^T \Sigma \tilde{B}^T \\ 0 \end{bmatrix} \begin{bmatrix} V^T & 0 \\ 0 & I_m \end{bmatrix}$$ \hspace{1cm} (4)

Let $R = \begin{bmatrix} \Sigma & \tilde{U} \tilde{\Sigma} \tilde{V}^T \\ 0 & \tilde{B}^T B \end{bmatrix}$, a square matrix of size $n+m$, the SVD: $R = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ can be efficiently computed regardless of $n$ [15]. Substituting the SVD of $R$ to Eq. (4), it is easy to obtain: $U' = [U \tilde{B}]U$, $\Sigma' = \tilde{\Sigma}$, $V' = \tilde{V}$.

Usually, we keep only $N_p$ eigenbasis ($N_p=1000$ or the number of samples) in the incremental learning process. After EBIF features are projected into the low-dimensional eigenspace obtained by incremental PCA, then Linear Discriminant Analysis can be efficiently performed.

## 5. Experimental Results

Three face datasets are use to evaluate the proposed EBIF descriptor: FERET [13], ORL [16] and the mobile face dataset gathered by ourselves. Typical samples of the three datasets are shown in Figure 2.

![Typical face samples of the three used datasets](image)

(a) FERET (b) ORL (c) our mobile face dataset

### 5.1. Experiments on the FERET dataset

The FERET dataset [13] is widely used in face recognition evaluation. It is divided into five subsets: a gallery set Fa consisting of 1196 subjects with one frontal image for each, the Fb subset containing 1195 face images with variations in expression, theFc subset containing 722 face images taken in an elapsed time with respect to the gallery set, and the DupI, a subset of Dup, containing 234 images in which the elapsed time is at least one year. In
FERET, there is also a training set with 1002 frontal images for model learning. As in [4], all the images were normalized to $128 \times 160$ by using the manually located centers of the eyes before EBIF feature extraction.

As comparison, results of both some classical algorithms and state-of-the-art ones are also given in Table 1. It can be seen that our EBIF method outperforms other state-of-the-art algorithms on the Fa, DupI and DupII subsets of FERET, and just a little inferior to HGPP [24] and HOGOM [4] on the Fc subset. The about 14% improvement over other state-of-the-art algorithms (e.g., HGPP[24], POEM[20], and HOGOM[4]) on the DupI subset indicates that EBIF method is more robust to aging faces (DupI images were taken in an elapsed time with respect to the gallery set).

We also investigate how different statistical features in the C1 layer of EBIF influence the recognition performance. As shown in Table 2, among the three single statistics: mean, std. dev. (standard deviation), and max, and their combinations, single max performs the worst, and the combination of “mean + std. dev.” performs best. Thus, in the C2 layer of EBIF, the combination of pyramidal statistics of mean and standard deviation is a better feature condensation approach than the max pooling operation.

| Table 1. Top rank recognition rates on the FERET dataset |
|-------------|---|---|---|---|
| methods     | Fb | Fc | DupI | DupII |
| Fisherface  | 94.0 | 73.0 | 55.0 | 31.0 |
| weighted LBP [1] | 97.0 | 79.0 | 66.0 | 64.0 |
| weighted LGBP[25] | 98.0 | 97.0 | 74.0 | 71.0 |
| HGPP [24] | 97.6 | 98.9 | 77.7 | 76.1 |
| POEM [20] | 97.6 | 96.0 | 77.8 | 76.5 |
| HOGOM [4] | 98.0 | 99.5 | 76.8 | 78.2 |
| EBIF        | 99.3 | 97.1 | 91.0 | 85.0 |

| Table 2. Top rank recognition rates of different statistical features in EBIF on the FERET dataset |
|-----------------|---|---|---|---|
| statistical features | Fb | Fc | DupI | DupII |
| mean            | 99.2 | 95.4 | 89.1 | 82.9 |
| std. dev.       | 98.7 | 95.9 | 89.6 | 85.0 |
| max             | 99.1 | 95.9 | 89.1 | 82.1 |
| mean + std. dev. | 99.3 | 97.1 | 91.0 | 85.0 |
| mean + max      | 99.3 | 96.3 | 90.2 | 84.2 |
| max + std. dev. | 99.1 | 96.9 | 88.9 | 83.8 |
| all three       | 99.2 | 96.9 | 89.8 | 83.3 |

5.2. Experiments on the ORL dataset

To gain knowledge about the robustness of our method against slight variations of pose angle and alignment we tested our approach on the ORL face database [16]. The ORL dataset [16] contains images from 40 individuals, each providing 10 different images. For some subjects, the images were taken at different times. There are variations in facial expression (open/closed eyes, smiling/non-smiling), facial details (glasses/no glasses) and scale (variation of up to about 10%). Moreover, the images were taken with a tolerance for some tilting and rotation of the face of up to 20 degrees. All images are grayscale and normalized to a resolution of $92 \times 112$ pixels.

As in [23], we use the first five image samples per class for training, and the remaining images for test. Thus, the total numbers of training samples and testing samples are both 200. The top rank recognition rates are shown in Table 3. It can be seen that the EBIF method performs much better than both the holistic methods (Eigenface, Fisherface and 2DPCA) and local appearance method LBP on this dataset. The good performance of EBIF on the ORL dataset is mainly because of its scale- and position-tolerant features obtained in the C1 layer.

| Table 3. Top rank recognition rates on the ORL dataset (50% for training and 50% for test) |
|-----------------|---|---|---|---|
|                | Eigenface | Fisherface | LBP | 2DPCA [23] | EBIF |
| Fb              | 91.5       | 94.5       | 96.5 | 96.0       | 97.5 |

5.3. Experiments on the mobile face dataset

Since there is no publicly available mobile face dataset for evaluation, we gathered one. Typical samples are shown in Figure 2 (c). All face images were captured by frontal cameras of iPhone/iPad or Android phones. The motivation of this dataset is to study face authentication on mobile devices. There are 78 subjects, and 439 face images for training and 242 for test respectively. As shown in Figure 2 (c), challenges of mobile face images include unevenly distributed illumination over face (light often comes from one direction), distortion (too close to the camera), pose changes, and etc. The ROCs of our EBIF method and three other classic methods (Eigenface, Fisherface, LBP) are shown in Figure 3. Obviously, Eigenface performs the worst, our EBIF method performs the best, and the second best is LBP. The Equal Error Rates (EER) of the four methods are 2.0% (EBIF), 3.5% (LBP), and 8.4% (Fisherface) and 15.1% (Eigenface) respectively.

6. Conclusion

This paper proposed an EBIF (Early Biologically Inspired Features) descriptor for face recognition, in which the features in the base model of BIM (Biologically Inspired Model) are improved by using more robust pyramidal statistics and applying L2 normalization in the C1 layer. The EBIF descriptor contains local normalized multi-orientation edge information and is slightly position- and scale-tolerant, thus showed good robustness to appearance variations caused by facial expressions, illumination, aging, pose and so on. To reduce irrelevant features in EBIF fea-
ture set, incremental PCA and LDA are combined to learn a discriminative subspace efficiently for feature dimension reduction. Experimental results on the public FERET and ORL datasets show that our EBIF method outperforms state-of-the-art face recognition algorithms. Also, results on our mobile face dataset proved its promising potentials for mobile authentication applications.

References