Local Visual Primitives (LVP) for Face Modelling and Recognition

Xin Meng[1], Shiguang Shan[2], Xilin Chen[2], Wen Gao[1,2]
[1] *School of Computer Science and Technology, Harbin Institute of Technology, China*
[2] *ICT-ISVISION FRJDL, Institute of Computing Technology, CAS, Beijing, China*
{Belinda, sgshan, xlchen, wgao}@jdl.ac.cn

Abstract

This paper proposes a novel simple yet effective generative model based on Local Visual Primitives (LVP) for face modeling and classification. The LVPs, as the pattern of local face region, are learnt by clustering a great number of local patches. Visually, these LVPs correspond to intuitive low-level micro visual structures very well, and they are expected to constitute those high-level semantic features, such as eyes, nose and mouth. We show that, though face appearances vary dramatically, these LVPs are very effective for face image reconstruction. For face recognition, block-based histograms of the LVPs indexes are extracted as the face representation to compare for classification. Primary experiments on FERET face database have shown that the LVP method can achieve encouraging recognition rate.

1. Introduction

Human face has been extensively studied for many years in computer vision and graphics for a wide range of missions such as detection, recognition, expression analysis, and animation. Especially, face recognition, as one of Biometrics, has been studied for more than three decades due to its wide potential applications in security and law enforcement. Appropriate face representation modeling the intrinsic face attributes is essentially important for all these face related tasks.

Among available face representation methods, some generative models, which can explicitly reconstruct the input face image using some model parameters, have attracted much attention. Eigenface [1], Active Shape Model (ASM) [2], Active Appearance Model (AAM) [3], 3D Morphable Model [4], Gabor Wavelet Net (GWN) [5], High Resolution Grammatical Model[6] are among them.

Eigenface [1] computes a set of orthogonal basis vectors, referred to as eigenvectors, from training face images by Principal Component Analysis (PCA). The linear combination of those eigenvectors can be utilized to describe and approach face images in the sense of least mean squared error. Thus, the coefficients of the linear combination can be regarded as the model parameters for reconstruction and classification.

ASM [2] is a statistical shape model based on Point Distribution Model (PDM), which represents the object shape based on the PCA analysis of the concatenated coordinates of some pre-defined landmarks. The model is then employed to constrain the landmark position obtained based on a local profile search. The object is finally represented as the parameters of the shape model.

AAM [3] further combines shape and texture to form the appearance model, which can synthesize a face image using hundreds of parameters. Then, given a novel face image, an analysis-by-synthesis iterative loop is exploited to estimate the model by minimizing the difference between current model and target image.

3D Morphable Model (3D MM) [4] have mad a further step to modeling both the 2D and 3D shape and texture by similarly using PCA as the statistical analysis tool. Moreover, in this model, graphical technologies are employed for rendering by introducing parameters for illumination, projection, and viewpoint. Given a face image, all the parameters are estimated by an optimization procedure.

In GWN [5], an object is described by the linear combination of a cluster of Gabor wavelets reflecting the local image structure. The parameters (position, orientation, and scale) of each Gabor wavelet are optimized by minimizing the difference between the model and the target.

More recently, the High Resolution Grammatical Model (HRGM) [6] has further extended the AAM with two additional layers. One layer uses a set of learned face components to refine the global AAM model. While the second layer represent skin marks and wrinkles with a learned dictionary of sketch primitives. It can achieve nearly lossless coding of face at high resolution.

Besides the above-mentioned generative models, Local Binary Pattern (LBP) has attracted more and more attention recently for texture analysis, face detection and recognition [7,8]. State-of-the-art results have been reported for face recognition. Due to its

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

gradient orientation nature, however, it is hard to extend to generative model for other face related task.

Inspired by basic ideas in GWN, HRGM and LBP, and motivated by the observations that both faces and facial components are structural patterns with micro-patterns, in this paper, we propose to model face images using the Local Visual Primitives (LVP) for both face modeling and classification. We show that LVP not only can generate the face images but also provide the basis for face recognition. Primary experiments have shown encouraging results.

2. Local Visual Primitives (LVP)

2.1 Motivation and Basic Ideas

In terms of perception, a face is composed of a few ***components***: forehead, two eyebrows, two eyes, nose, cheek, mouth, and chin. Among them, the forehead, the cheek, and the chin are by and large plain regions with little variation, while the eyes, the nose, and the mouth can be further decomposed structurally to some ***sub-parts***. For instance, an eye is composed of the upper and lower eyelids, the eyelash, the white, and the iris. It seems that, from the viewpoint of perception, the decomposition ends here. However, we argue that, from the angle of computation, a further step can be made forward. Just like the morpheme in syntax-language exists in the mid-level of character and word, we argue that there should be a mid-level between the sub-part and pixel. As illustrated in Fig.1, before coming to the end of ***pixels***, we argue that the sub-parts can be further decomposed into smaller basic units, called by us as Local Visual Primitives (***LVPs***).

An LVP is actually the structural grouping of some pixels that appear together more frequently in perception in terms of probability. In some sense, they can be regarded as something like a "larger pixels". Though these primitives may not correspond to any intuitive perceptional concepts, they may provide a better basis for image formation than the isolated pixels. The image representation based on LVPs should be more compact and more robust to variations than isolated pixel-based image representation, because LVPs actually perform like some kind of ***local descriptors*** modeling the facial microstructure.

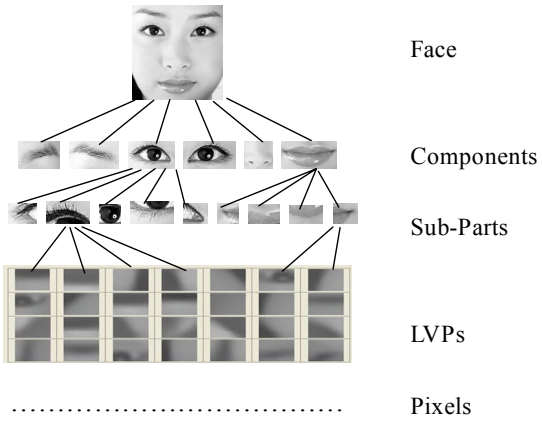

Figure 1. Illustration of LVP-based Face Model

2.2. Learn LVP Dictionary for Faces

At first thought, one may feel that it is almost impossible to find the entire LVPs set for face images because of the immense population on the earth. However, faces are similar anyway, and the sub-parts of faces look more similar. Therefore, it is possible to learn an LVP dictionary for faces. To verify this, we have conducted primary experiments on the seeking of the LVPs for faces by using clustering method.

Since faces are structural objects that can be aligned coarsely in terms of scale, rotation, and translation according to the two eyes, we can learn LVPs at a fixed scale and ignore the rotation and translation, if only we align all the face images in the same mode according to the two eye positions. Therefore, in this primary contribution, we only consider the simplest case where the LVPs are a rectangle image patch with fixed size, e.g. *m* by *m* pixels. Clustering algorithm is then employed to learn those most representative patches to form the LVP dictionary for face modeling.

In our implementation, given a training set with sufficient face images aligned according to the eyes, we densely sample plenty of *mxm* rectangle patches from the training images. In addition, for each image patch, the mean intensity is subtracted from the intensity of each pixel to remove the DC component, since we care only the local variation rather than the absolute intensity. Then a clustering algorithm is conducted on these image patches to obtain *N* clusters. The mean of each cluster is defined as the Local Visual Primitives. Thus, *N* LVPs is obtained and the LVP dictionary is denoted as:

$$\Omega_{\mathrm{LVP}} = \{w_1, w_2, \cdots, w_N\} \quad (1)$$

Fig.2 illustrated some examples of the LVPs with *m*=11 and *N*=256, we learnt from a training set containing images from the FERET face database.

Figure 2. LVPs learned from FERET faces

From Fig.2, we can see these LVPs also correspond to intuitive low-level micro-image structure such as edges, circles, etc.

3. Face Reconstruction Based on LVPs

As local microstructures, LVPs can be used to reconstruct face images. The basic idea is that we can seek for the most similar LVP for the image patch at each image position or at a certain sampling step, and then replace the image patch by the LVP. Note that for each image position, its reconstructed intensity may be the weighted mean of corresponding pixels from several overlapped LVPs. Fig.3 shows an example reconstructing a face image incrementally using more and more LVPs.

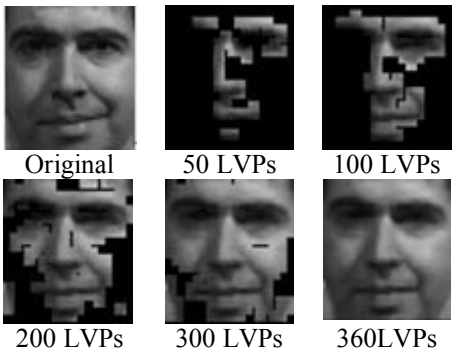

Figure 3. An example of face reconstruction using some amount of LVPs. The LVP size for this example is 5x5 and the face size is 56x63.

The algorithm reconstructing a face image is as follows:

(1) Sample the input image at a sampling step (e.g. 2 pixels) to obtain N sub-images of the same size as LVPs, and denote them as : $\{B_i : i = 1,...,N\}$;

(2) For each sub-image B_i , compute a weight $\varphi_i = E_i + S_i$, where E_i is the total edge energy of B_i , and S_i is the similarity between B_i and the LVP ω_i most similarity to B_i . Then, all the weights are sorted in descending order.

(3) Initialize the intensity of each pixel of the reconstructed image to be 0. Then, for j=1 to M (M is the predefined total number of LVP).

a) Take the j-th largest weight φ_j ;

b) Additionally update the intensity of pixels in B_j by the corresponding LVP w_j .

Note that the edge energy E_i is exploited in step 2 in order that the most salient facial regions, such as the eyes, the nose, and the mouth, can be first reconstructed. As can be seen from the example in Fig.3, only hundreds of LVPs can effectively reconstruct an input face images reasonably well.

4. LVPs for Face Recognition

LVP actually can be regarded as a kind of local descriptor, which can extract the local features. Local descriptors, such as Gabor, Local Binary Pattern, have attracted more and more attention in recent years for their robustness to local distortions. We then propose the following LVP-based method for face recognition. The method is composed of two stages: extraction of face template and template matching.

4.1 Extraction of Face Template

In our method, an input face image is finally modeled as a set of histograms by the following procedure:

(1) First, at each image position (x, y), compute the index of the LVP most similar to the m*m sub-image centering at (x, y). And denote this LVP index as $w(x, y)$. Thus, an LVP index "map", denoted as Δ_{LVP}, of the same size as the original image can be obtained.

(2) Then, Δ_{LVP} is spatially partitioned into K blocks of some specific size (α, β) . We denote these blocks as $\{R_i : i = 1, \cdots, K\}$.

(3) Next, in each block R_i , we compute a histogram of the LVP index and denote as H_i . Thus, the feature template of the input face image is finally represented as a set of histograms:

$$\Re_{LVP} = (H_1, H_2, \cdots, H_K) \qquad (2)$$

Easy to see, in the above procedure, LVP is actually used as an operator assigning each image position a LVP mode that most matches its local neighborhood.

4.2 Face Template Matching

Given two input images, face recognition essentially needs to compute their similarity. After the feature templates are extracted using Equ.2, their similarity is computed using the following equation:

$$S(\Re, \Re') = \sum_{i=1}^{K} S_{HI}(H_i, H_i') \qquad (3)$$

where

$$S_{HI}(H_i, H_i') = \sum_{b=1}^{L} \min(h_i(b), h_i'(b)) \qquad (4)$$

is the histogram intersection commonly used to match two histograms, in which L is the number of bins for each histogram. $S(\Re, \Re')$ is then used for face recognition based on the Nearest Neighbor classifier.

5. Experiments for Face Recognition

Some primary experiments are conducted on the *fafb* probe set in the standard FERET face database. In our experiments, we evaluate the method based on the standard gallery (1196 images of 1196 subjects) and the fafb probe sets (1195 images). Please refer to [9] for details about the FERET database.

The LVP dictionary is learned from a subset of the FERET training CD. The subset contains 1002 frontal images of 429 subjects. With the size of the normalized face being 56 by 63 pixels, we have also tried varying configuration of the method, including the LVP size (m*m), the size (N) of the LVP dictionary, and the number (K) of blocks (or the block size given face image size). Table 1 shows the rank-1 recognition rates of these configurations. In the table, we also give the known best results on the same *fafb* probe set, which is published on ICCV2005 [8].

From the table, we can observe that the performance of the proposed method is very comparable with the known best results. Another advantage of the method is that its recognition rate does not change dramatically with varying configurations, as can be seen from table 1, which can greatly facilitate the system design.

Table 1. Experimental results of the proposed method with varying configurations

Configuration of the Method			Accuracy on ***fafb*** probe set
LVP size (mxm)	#LVP (N)	#Blocks (K)	
8x8	256	72	93.1%
5x5	256	49	96.2%
5x5	256	12	94.0%
5x5	512	49	96.7%
5x5	256	126	96.2%
3x3	256	12	93.8%
3x3	256	72	97.5%
3x3	256	126	97.2%
Known best reported results [8]			98.0%

6. Conclusion and Discussions

This paper proposes Local Visual Primitives (LVP) for face modeling and classification. The LVPs are learned by clustering a great number of local patches. Visually, these LVPs correspond to intuitive low-level micro visual structures very well. And they can efficiently reconstruct face images, though face appearances vary dramatically. For face recognition, block-based histograms of the LVPs indexes are extracted as the face representation to compare for classification. Primary experiments on FERET face database have shown that the LVP method can achieve encouraging recognition rate.

LVP may seem somewhat similar to LBP. However, compared with LBP which can only modeling the ordinal relationship between neighboring pixels, the proposed LVP can model more information including the intensity difference. And LVP is also applicable to image reconstruction, which is impossible for LBP.

Acknowledgement

This research is partially sponsored by Natural Science Foundation of China under contract No.60332010, and the Program for New Century Excellent Talents in University (NCET-04-0320).

Reference

[1] M.Turk, A.Pentland, “Eigenfaces for recognition. J. of Cognitive Neuroscience”, vol.3, pp. 71-86, 1991

[2] T.F.Cootes, C.J.Taylor, D.H.Cooper and J.Graham, “Active Shape Models – Their Training and Application”, Computer Vision and Image Understanding, 61(1), pp. 38-59,1995

[3] T.F.Cootes, G.J.Edwards, C.J.Taylor, “Active Appearance Models”, Proc. European Conf. Computer Vision, vol. 2, pp. 484-498, 1998

[4] V.Blanz, T.Vetter, “Face Recognition Based on Fitting a 3D Morphable Model”,TPAMI,25(9), 1063-1075, 2003

[5] V. Kr¨uger. “Gabor wavelet networks for object representation” Technical Report CS-TR-4245, University of Maryland, CFAR, May 2001

[6] Z.J.Xu, H Chen, S.C. Zhu "A High Resolution Grammatical Model for Face Representation and Sketching". In Proc. of CVPR2005, pp. 470-477

[7] T.Ojala, A.Hadid, M.Pietikäinen. “Face recognition with Local Binary Patterns”, ECCV, pp.469-481, 2004.

[8] W.Zhang, S.Shan, W. Gao, and X. Chen, “Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition”, ICCV’05, pp. 786-791, 2005.10

[9] P.J. Phillips, H.M. Syed, A. Rizvi, and P.J. Rauss. “The FERET evaluation methodology for face-recognition algorithms”. PAMI, 22(10), pp. 1090-1104.