ABSTRACT

Traditional appearance-based head pose estimation methods use the holistic face appearance as input and then employ statistic learning methods to extract low dimension features for classification. However, the face appearance may be more related to the unique identity of an individual rather than head poses. In this paper, we propose an image descriptor, covariance of oriented gradients (COG), for head pose estimation. This descriptor computes the covariance matrix of gradient based image features which characterizes the geometry structure of head pose images. To incorporate the spatial information, a head image is divided to several cells and the covariance matrices of all the cells are combined as the original image descriptor. Under the Log-Euclidean metric, the original image descriptor is mapped into vector space, and then linear discriminant analysis is employed to find the discriminative low dimensional features. Experiments show that the proposed method outperforms two other state-of-the-art methods in terms of estimation accuracy and robustness on image resolutions.

Index Terms—Head pose estimation, covariance of oriented gradients, Log-Euclidean metric

1. INTRODUCTION

Head pose is an important cue in many computer vision and human computer interaction applications [1]. Among the different head pose estimation methods, appearance-based methods have attracted significant attention in recent years, in particular due to their suitability for use on low resolution images. Appearance-based methods formulate the head pose estimation as a pattern classification problem. Previous work has used different statistical learning methods for head pose estimation, including PCA[2], ICA [3], kernel PCA and kernel LDA [4], SVM [5] etc. Tu et al. [6] proposed to locate the nose-tip and estimate head pose using a Tensorposes model. In recent years, manifold learning methods have also been employed to learn the intrinsic low dimensional structure of head pose data [7]. Refer to [1] for a detailed survey on head pose estimation methods.

One of the key problems in appearance-based methods is to find proper features which can capture the discriminative information of pose variations and are insensitive to other variations like identity, lighting etc. Traditional appearance-based methods typically use holistic face appearance as input and then employ statistical learning methods to extract low dimensional features for classification. However, face appearances contain not only the pose information but also information about identity, lighting etc. which limits the performance of the appearance-based methods. Researchers have proposed some other features which emphasize the pose variations and suppress other variations. A Laplacian-of-Gaussian filter [7] was used on the images to emphasize the facial contours while removing the identity-specific texture variations. Gabor wavelets have been explored in [8]. Histogram of oriented gradient (HOG) descriptor [9], which can describe the distribution of gradient orientations, has been used for head pose estimation [10]. Ma et al [11] proposed a GaFour method that exploits the asymmetry of facial appearance. They show that the GaFour feature combined with LDA outperforms other methods such as PCA, LDA, ICA and the Gabor Fisher Classifier [12].

Based on the assumption that head pose is more related to the geometry structure of face images rather than the appearance, we investigate how to represent the geometry of face images in low resolution images. The application of HOG descriptor in head pose estimation shows that gradient features are suitable for characterizing the head pose variations while they are insensitive to other facial variations like identity and lighting. We try other ways to combine the gradient features into a compact representation which is more discriminative between different head poses.

Tu et al. [13] proposed a covariance matrix descriptor to characterize the appearance of an image region. This descriptor is capable of capturing the variances of each feature and the correlations between different features inside an image region. The covariance matrix descriptor is insensitive to lighting, view and pose variations, and it has been successfully applied in object detection [13], texture classification [13] and object tracking [14]. The covariance matrix is a symmetric positive definite (SPD) matrix and it lies on a Riemannian manifold. Statistics for covariance matrices is usually computed through Riemannian geometry based on the affine-invariant metric [13], under which computing the distance between two SPD matrices needs solving the generalized eigenvalues of the two matrices. Recently, a new Log-Euclidean metric is proposed [15].
Under this metric, the distance between two SPD matrices takes a much simpler form which is similar as the Euclidean distance. Therefore, the covariance matrices can be mapped to vector space and then machine learning methods in vector space can be exploited. In [16], after mapping the covariance matrices to vector space under the Log-Euclidean metric, incremental PCA is used to learn the low dimensional eigenspace model for object tracking.

At first sight, one might think that the covariance matrix descriptor is not suitable for characterizing the discriminative information of head poses due to its insensitivity to view and pose variations. However, in fact, the covariance matrix has very strong descriptive capability so that it can capture the differences between various head poses. Therefore, in this paper, we propose to use the covariance matrix of gradient based image features for characterizing the geometry structure of head pose images, and we call the image descriptor as covariance of oriented gradients (COG). To incorporate the spatial information, an image is divided to a grid of equal-sized cells and the COG of each cell is computed separately. The Log-Euclidean metric is employed to map the covariance matrices to vector space, and then linear discriminant analysis is used on the mapped vectors to obtain low dimensional discriminative features. Experimental results show that the proposed method outperforms two other state-of-the-art methods in terms of estimation accuracy and robustness on image resolutions.

2. COVARIANCE OF ORIENTED GRADIENTS

Tuzel et al. [13] proposed a covariance matrix descriptor to characterize the appearance of an image region. Let $I$ be a $W \times H$ one dimensional intensity or three dimensional color image. Denote $F$ as the $W \times H \times d$ dimensional feature image extracted from $I$:

$$F(x, y) = \phi(I, x, y),$$

where $\phi$ is a mapping function for extracting the image features such as intensity, color, gradients, and filter responses, etc. For a given rectangular region $R \subseteq I$, let $\{f_i\}_{i=1}^{N}$ be the $d$-dimensional feature points obtained by $\phi$ within $R$. Then the image region $R$ can be represented as a $d \times d$ covariance matrix:

$$C_R = \frac{1}{N-1} \sum_{i=1}^{N} (f_i - \mu)(f_i - \mu)^T,$$

where $\mu$ is the mean of the feature points.

Based on the assumption that head pose is more related to the geometry structure of face images rather than the appearance, to characterize the pose information of a head image, we define the mapping function $\phi(I, x, y)$ as

$$\begin{bmatrix} x & y & I_x & I_y & I_{xx} & I_{yy} & \sqrt{I_x^2 + I_y^2} \arctan(I_x/I_y) \end{bmatrix}^T,$$

where $x$ and $y$ are the pixel locations, $I$ is the image intensity, $I_x$, $I_{xx}$, $I_y$, $I_{yy}$ are the first and second order intensity gradients in both directions, $\sqrt{I_x^2 + I_y^2}$ is the magnitude of the first order intensity gradient, and the last term is the edge orientation. With the defined mapping function, each pixel of the input image is mapped to a $d=9$ dimensional feature. Thus the covariance descriptor of a region is a $9 \times 9$ matrix, which has only 45 different values.

Since our mapping function is mainly related to the gradient features, we call this descriptor as covariance of oriented gradients (COG), as an analogue to histogram of oriented gradients (HOG). The COG descriptor encodes the information of the variances of the gradient features, their correlations with each other and the spatial layout of these features. Although pixel locations are included in the feature mapping function, the COG descriptor still loses some global spatial information which might be very important for discriminating head poses. To incorporate the spatial information, we divide the original image to $M \times N$ cells and compute the COG for each cell separately. The set of all the COGs makes up the original image representation for a head image. In our study, we set $M=N=4$.

3. THE LOG-EUCLIDEAN METRIC AND IMAGE DESCRIPTOR IN VECTOR SPACE

SPD matrices lie on a connected Riemannian manifold. Therefore, Riemannian metrics should be used for statistics on SPD matrices. Recently, Arsigny et al. proposed a Log-Euclidean Riemannian metric [15] under which there is a mapping relationship between the Riemannian manifold and the vector space of SPD matrices. Under the Log-Euclidean metric, the distance between two SPD matrices $X$ and $Y$ is calculated by $\| \log(Y) - \log(X) \|$. Clearly, the distance takes a much simpler form and can be computed easily in a vector space structure.

Under the Log-Euclidean metric, our proposed image descriptor in vector space is computed as follows. Let a covariance matrix $C_{m,n}$ be the COG descriptor of cell $(m,n)$. By the Log-Euclidean mapping, $C_{m,n}$ is transformed to the matrix logarithm $\log(C_{m,n})$. Due to its vector space structure, we unfold $\log(C_{m,n})$ into a vector. Since a covariance matrix and its matrix logarithm are both symmetric, there are only $d \times (d+1)/2$ independent values. We define the vector operation of a symmetric matrix $C$ as

$$\text{vec}(C) = \begin{bmatrix} c_{1,1} & \sqrt{2}c_{1,2} & \sqrt{2}c_{1,3} & \ldots & \sqrt{2}c_{2,2} & \ldots & c_{d,d} \end{bmatrix}^T.$$
By multiplying the non-diagonal elements by $\sqrt{2}$, the distance between the two feature vectors will be the same as that between the two covariance matrices under the Log-Euclidean metric. So the unfolded vector from $\log(C_m,n)$ will be $\text{vec}(\log(C_m,n))$. Finally, the unfolded vectors of all the cells are concatenated to make up a long feature vector whose dimension is $M \times N \times d \times (d+1)/2$. The steps of computing the final image descriptor is illustrated in Fig. 1.

Fig. 1. Flow chart of the COG descriptor

4. EXPERIMENTS

In this section, we evaluate the performance of our method by comparing it with three other image descriptors: Image intensity, HOG [10] and GaFour [11]. All the descriptors are reshaped to 1-D vector as the original image feature. We first employ the PCA for dimension reduction, and then utilize the LDA to find an optimal low dimensional subspace so that the samples within class can be clustered together and the samples between classes can be separated as much as possible. The original high dimensional image feature is transformed by PCA and LDA to generate low dimensional features for classification. In this study, when PCA was used for dimension reduction, typically 95% of the total energy is kept. For the HOG and COG feature, we set the ratio at 99.9% with which the best accuracy can be obtained. After LDA, the dimension can be reduced to $c - 1$, where $c$ is the number of the class labels. Given a face image, after PCA and LDA, the low dimensional feature is classified to a predefined class label. In this study, we use the nearest centroid (NC) classifier. We used a multi-view face detection method to locate the face bounding box from the images and the cropped face images were resized to 32×32. Finally the face images are normalized with zero mean and unit variance to reduce the influence of lighting.

We evaluate the performance on a public CAS-PEAL database [17], in which each individual has 21 poses with seven yaw angles (-45° to 45° with 15° step) and three pitch angles (-30°, 0°, and 30°). In this study, we used a subset of 200 subjects whose IDs ranged from 401 to 600 and there are 4200 images in total. All the images with the same yaw angles are grouped under the same class label yielding 7 classes with 600 images in each class. So for all the LDA based methods, the dimension of features will be 6. We split the data into 4 equal subsets and performed 4-fold cross-validation. Each subset has 50 subjects with a total of 50×21=1050 images. In each run, 3 subsets are used as the training set and the other subset is used as a test set. The subjects in the training and test set are completely distinct. All the results reported are the average results of 4 runs.

Table 1. Accuracy (%) of different methods

<table>
<thead>
<tr>
<th></th>
<th>PCA+</th>
<th>Image</th>
<th>GaFour</th>
<th>HOG</th>
<th>COG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>66.33</td>
<td>77.43</td>
<td>76.36</td>
<td>84.71</td>
<td></td>
</tr>
<tr>
<td>Face</td>
<td>82.38</td>
<td>91.29</td>
<td>91.86</td>
<td>95.33</td>
<td></td>
</tr>
<tr>
<td>Face+up sampled</td>
<td>83.83</td>
<td>91.29</td>
<td>91.86</td>
<td>95.33</td>
<td></td>
</tr>
</tbody>
</table>

For the NC classifier, using only one centroid may be not sufficient to represent the samples of each class. So for each pose angle, we used the $k$-means method to find the $k$ centroids from the training samples. Figure 3 shows the accuracies for various values of $k$ from 1 to 10. From the figure, we can see that the accuracy of COG+LDA is best whatever the value of $k$. For the supervised method LDA, the accuracies are stable for different values of $k$, which suggests that in the subspace obtained by LDA, the samples of each class are clustered together very well.

To evaluate the robustness of the methods to image resolutions, we down sampled the face region from the original image to 8×8, 16×16, 32×32, 64×64 and we repeated the experiments for each size. And for 8×8, 16×16 images, we also up sampled them to 32×32 and repeated the experiments. Figure 4 shows the accuracy under different image sizes. From the figure, we can see that the accuracy of COG+LDA is always the highest for all the sizes, which demonstrates its superior performance and robustness. Even in the 8×8 image, the accuracy is as high as 86.02% and after up sampling, the accuracy increases significantly by 6.29% to 92.31%. It can also be found that the unsupervised method COG+PCA is always better than the supervised method Image+LDA. This proves that the COG descriptor can describe the pose variations very well, while image intensity is more sensitive to other variations.
We evaluated the performance when dividing an image to different number of cells. The result is listed in Table 2. From the table, we can see that if there is only one cell, the accuracy is the worst. If the image is divided to 2x2 cells, the accuracy increases significantly. When the image is divided to more cells, the accuracy will increase gradually. This proves that the spatial division can really increase the discriminative capability of the image representation.

The average time for computing the original COG feature is around 10.8 ms and the average time for computing the low-dimensional feature and classification is around 0.4 ms in Matlab 7.0.4 on a PC (CPU: Pentium IV 2.80-Ghz, RAM: 2-GB). This indicates that our method can be used in a real-time system.

Table 2. Accuracy (%) of different cell division methods

<table>
<thead>
<tr>
<th>Cells</th>
<th>1x1</th>
<th>2x2</th>
<th>3x3</th>
<th>4x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>81.29</td>
<td>94.43</td>
<td>94.55</td>
<td>95.33</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

We have proposed an image descriptor, covariance of oriented gradients with spatial layout division, for head pose estimation. Under the Log-Euclidean metric, the original image descriptor is mapped to vector space, and then LDA is used to find low dimensional discriminative features. Experiments show that the proposed method outperforms other state-of-the-art methods in terms of estimation accuracy and robustness on image resolutions.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under grants Nos. 60673189, 60873266 and 90820304. The authors would like to thank B. Ma for providing the GaFour feature extraction code.

6. REFERENCES