Occlusion, Expression and Illumination Invariant Face Recognition Using Block-based Bag of Words

Zisheng Li Jun-ichi Imai Masahide Kaneko

Graduate School of Electro-Communications, The University of Electro-Communications

1. Introduction

Face recognition has become one of the most active research fields due to the wide range of commercial and law enforcement applications. In this paper, we propose a novel block-based bag of words (BBoW) method for face recognition, which is robust to variant facial expressions, illumination, and occlusions, only using a single neutral expression frame per person for training. Experimental results show that our method significantly outperforms other recent works.

2. Block-based Bag of Words

Recently, the bag of words (BoW) method [1], which represents an image as an orderless collection of local features, has demonstrated impressive performance for object recognition. However, in face recognition, the object images belong to the same category (face images) and thus histograms of orderless local features from the whole face luck large enough between-class variations.

In this paper, we propose a block-based bag of words (BBoW) method to extract more subtle facial features for face recognition. As shown in Fig. 1, we partition the face image into 5×5 blocks, and consider each block as a ROI (region of interest). For each ROI, we calculate dense SIFT features on a regular grid with spacing 2 pixels. As a result, a set of dense SIFT features are obtained for each block of each face image. At the training stage, we convert the SIFT vectors of each ROI to different codewords using k-means algorithm respectively. The SIFT features of different grids from an ROI are partitioned into K clusters. A codeword is defined as the center of a learned cluster, and it is considered as the representative of the cluster of SIFT vectors from the block over the training set. A codebook consists of the K codewords of the same ROI, and there are totally 5×5 codebooks generated from the training data. As a result, each SIFT vector of a sliding grid in an ROI of a face image is mapped to a certain codeword in the corresponding codebook. And this ROI can be represented by the counting number of different codewords using histograms. Then, 5×5 histograms are concatenated together to represent the face image. Linear SVMs are applied to train the histograms of different subjects finally. At the testing stage, input face images are converted to 5×5 histograms of codewords using the trained codebooks, and concatenated to a whole one. Classification results can be obtained using the trained models by the SVM classifiers.

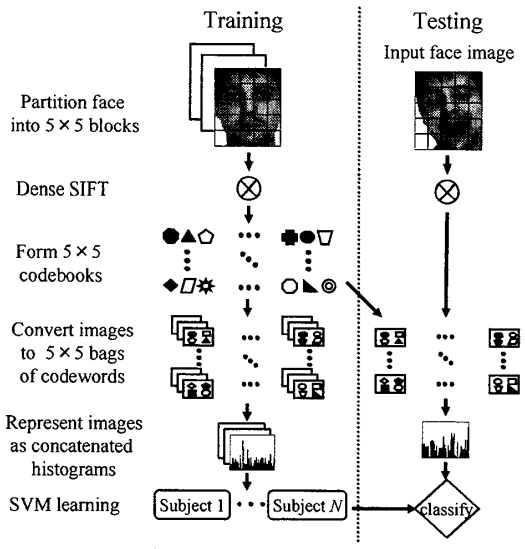

Fig. 1 Framework of BBoW

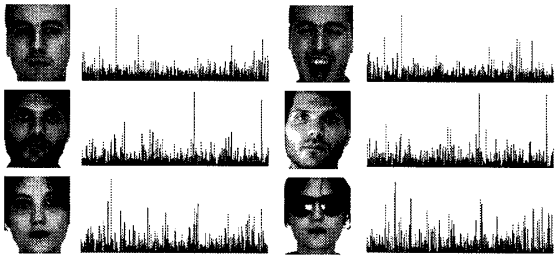

Fig. 2 Examples of BBoW features

Figure 2 shows some examples of BBoW features extracted from different subjects with different expressions, illuminations, and occlusions. We can see that BBoW features of the same subject in variant extreme conditions can still have similarities, and those of different subjects obviously have discriminative differences.

3. AR and XM2VTS Database

We use the AR database [2] and XM2VTS database [8] in our work. The AR database consists of over 3200

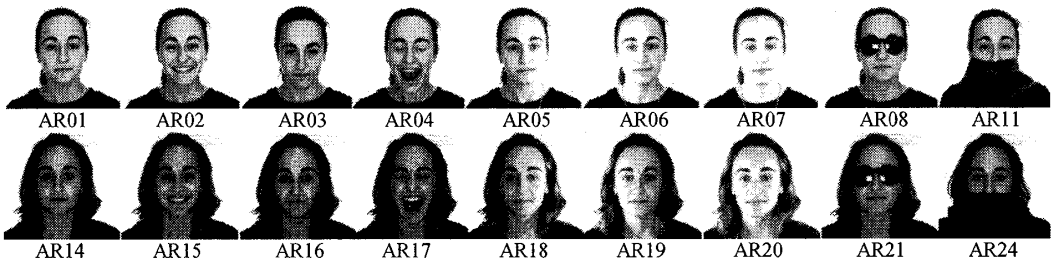

Fig. 3 Images of AR database (First row: first session; second row: second session)

Table I Robust face recognition results on AR database

Method	Subject Num	HE	Recognition results (%)															
			Facial expressions						Illuminations						Occlusions			
			AR02	AR03	AR04	AR15	AR16	AR17	AR05	AR06	AR07	AR18	AR19	AR20	AR08	AR11	AR21	AR24
[3]	50	-	94	96	57	-	-	-	-	-	-	-	-	-	80	82	49	52
[4]	117	-	96.58	87.18	38.46	-	-	-	74.36	64.96	42.74	-	-	-	70	72	58	60
[5]	100	Yes	100	98	88	88	90	64	-	-	-	-	-	-	97	95	60	52
[6]	121	Yes	98.8		-	-	-	-	98.9		-	-	-	-	-	-	-	-
[7]	112	-	78.57	92.86	31.25	-	-	-	92.86	91.07	74.11	-	-	-	-	-	-	-
Ours	**119**	**No**	**100**	**100**	**95.80**	**97.48**	**97.48**	**77.31**	**100**	**100**	**98.32**	**99.16**	**93.28**	**80.67**	**94.96**	**99.16**	**77.31**	**89.92**

face images of 126 subjects with variant facial expressions, lighting, and partial occlusions. There are 26 different images per person, recorded in two different sessions separated by two weeks. 119 out of 126 subjects have complete sets of images covering all conditions. Figure 3 shows images of one subject in both sessions.

The XM2VTS database consists of 2360 face images from 295 subjects varying in poses, hairstyles, expressions and glasses. There are eight shots of each person, obtained from four sessions spread out in monthly intervals.

4 . Experimental Results

Face images are cropped to 230×270 pixels without histogram equalization or any further image alignment. According to a series of overall recognition experiments, we choose a sampling grid of 4×3 pixels to calculate SIFT features, the sampling interval is 2 pixels. The codebook size is $K = 75$.

We firstly use images AR01~AR07 of the 119 subjects as training data, AR14~AR20 as testing data to test the overall performance of our method. The average recognition rate is 98.92%, which is the best result ever using the above protocol. For the XM2VTS database, we use images of 295 subjects in the first three sessions as training data, and the rest as testing data. The average recognition rate reaches 100%.

In order to test the robustness of our method, we train our system only using a single frame with neutral expression per subject, AR01 in the AR database, and test the rest images as shown in Fig. 3. Table I shows comparison results with previous works. Our method significantly outperforms other systems in all conditions except that in AR08, [5]'s result is better than ours. It should be noted that [5]'s work assumed that the occlusions were known and removed from the testing images beforehand. Moreover, in [5], both training and testing images were preprocessed by histogram equalization (HE) for images with different expressions, but for images with occlusions, HE was not processed to avoid unwanted effect. On the other hand, our method uses the same parameters for all conditions without any illumination compensation or image alignment. It can be said that our method has the best performance ever for the AR database since it is able to deal with the largest number of variant conditions and it outperforms other recent works significantly.

5 . Conclusions

We propose a block-based bag of words method for robust face recognition. This is the first time to apply the bag of words method to face recognition. The proposed method can robustly give excellent results under various conditions including extreme expressions, strong non-uniform lighting and occlusions, using a single frame per person for training. Experimental results show that our method significantly outperforms other recent works.

Reference

[1] F. Li *et al.*, "A bayesian hierarchical model for learning natural scene categories," *Proc. of CVPR'05*, 2:524-531, 2005.
[2] A. M. Martinez and R. Benavente, "The AR face database," *CVC Tech. Rep.*, #24, 1998.
[3] A. M. Martinez, "Recognition imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(6):748–763, 2002.
[4] H. R. Kanan *et al.*, "Face recognition using adaptively weighted patch PZM array from a single exemplar image per person," *Pattern Recognit.*, 41(12):3799-3812, 2008.
[5] X. Tan *et al.*, "Recognizing partially occluded, expression variant faces from single training image per person with SOM and soft k-NN ensemble," *IEEE Trans. Neural Netw.*, 16(4):875–886, 2005.
[6] X. Xie and K. Lam, "Gabor-based kernel PCA with doubly nonlinear mapping for face recognition with a single face image," *IEEE Trans. Image Process.*, 15(9):2481–2492, 2006.
[7] Y. Gao and M. K.H. Leung, "Face recognition using line edge map," *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(6):764–779, 2002.
[8] K. Messer *et al.*, "XM2VTSDB: The extended m2vts database," *AVBPA*, 1999.