Using the Visual Words based on Affine-SIFT Descriptors for Face Recognition

Yu-Shan Wu, Heng-Sung Liu, Gwo-Hwa Ju, Ting-Wei Lee, Yen-Lin Chiu
Business Customer Solutions Lab.,
Chunghwa Telecommunication Laboratories
12, Lane 551, Min-Tsu Road Sec.5
Yang-Mei, Taoyuan, Taiwan 32601, R.O.C.
{yushanwu, lhs306, jgh, finas, lewis32330}@cht.com.tw

Abstract—Video-based face recognition has drawn a lot of
attention in recent years. On the other hand, Bag-of-visual
Words (BoWs) representation has been successfully applied in
image retrieval and object recognition recently. In this paper, a
video-based face recognition approach which uses visual words
is proposed. In classic visual words, Scale Invariant Feature
Transform (SIFT) descriptors of an image are firstly extracted
on interest points detected by difference of Gaussian (DoG),
then k-means-based visual vocabulary generation is applied to
replace these descriptors with the indexes of the closet visual
words. However, in facial images, SIFT descriptors are not
good enough due to facial pose distortion, facial expression and
lighting condition variation. In this paper, we use Affine-SIFT
(ASIFT) descriptors as facial image representation.
Experimental results on UCSD/Honda Video Database and
VidTIMIT Video Database suggest that visual words based on
Affine-SIFT descriptors can achieve lower error rates in face
recognition task.

Keywords-component; face recognition, SIFT, Affine-SIFT,
visual words.

I. INTRODUCTION

Video-based face recognition has been a popular research
topic because of its scientific challenges and wide use of
video monitoring. However, there are many well-known
approaches proposed to overcome the face recognition
problems. Among them, Principle Component Analysis
(PCA) [1] searches a subspace in the feature space that has
the largest variance and then projects the feature vectors onto
it. Linear Discriminant Analysis (LDA) [2] attempts to obtain
another subspace which can maximize the ratio of between-
class variance to the within-class variance. Locality
Preserving Projection (LPP) [3] also tries to find an optimal
linear transform that preserves local neighbor information of
data set in a certain sense.

Recently, methods based on multiple images/video
sequences for face recognition are also proposed. Mutual
Subspace Method (MSM) [4] considers the minimum angle
between input and reference subspaces as measure of
similarity, and each subspace is formed by PCA operation on
image sequence from each person. Constrained Mutual
Subspace Method (CMSM) [5][6][7] is an improved version of
MSM. The construction of input and reference subspaces are
the same as in MSM, except the bases of these subspaces are
further projected onto a constrained subspace and the
projected bases are used to calculate the similarity between
two persons.

All the above methods are concentrating on the
projection or transformation of feature vector. The feature
vector of a face image used by these methods is usually
simple gray value in row-major order. However, the feature
selection and extraction are also extremely important in face
recognition. Recently, Bag-of-visual Words (BoWs) [8][9]
image representation has been utilized in many computer
vision problems and has demonstrated impressive
performance. In this method, Scale-Invariant Feature
Transform (SIFT) [10] features of an image are firstly
extracted on interest points which are usually detected by
difference of Gaussian (DoG) method. Then a clustering
method is used to convert these SIFT features to codeword
histogram. Finally the degree of similarity between two
images can thus be measured by the distance between their
histograms.

Different face images of the same person obtained by
camera in varying position and angle undergo apparent
deformations. These deformations can be alleviated by affine
transform of image plane. The parameters of the affine
transform can be described scale, rotation, translation,
camera latitude and longitude angles. Although SIFT method
is invariant to three out of the above five parameters, it is not
good enough. ASIFT method [11] is proposed to cover all
five parameters and has been proved to be fully affine
invariant. Furthermore, the computation complexity of the
ASIFT method can be reduced to about twice of SIFT
method by a two-resolution scheme. In this paper, we
applied ASIFT Visual Words as face image representation.
Experimental results on UCSD/Honda [12] Video Database
and VidTIMIT [13] Video Database show that ASIFT Visual
Words method is superior to other classical methods.

This paper is organized as follows. In Section II, we
introduce three representations for face image, SIFT Method,
ASIFT method and the proposed method. In Section III,
experimental results based on famous UCSD/Honda Video
Database and VidTIMIT Video Database are depicted.
Finally, the discussion and conclusion are presented in
section IV.

II. FACE RECOGNITION APPROACHES

In this section, we introduce two image representations
mentioned previously in Section I and the proposed method.
In section 2.1 and 2.2, SIFT method and ASIFT method for face image representation are introduced respectively. In section 2.3, the proposed face recognition method is derived. Finally, the performance evaluation of face recognition in video sequences is introduced in section 2.4.

### 2.1 Scale Invariant Feature Transform (SIFT)

SIFT method compares two images by a rotation, a translation and a scale change to decide whether one image can be deduced from the other image. To achieve scale invariance, SIFT simulates the zoom in scale space. This can be accomplished by searching for stable points across all possible scales and these stable points can be thought to be invariant to scale change. The scale space of an image is formed by the convolution of this image with a variable-scale Gaussian $G(x, y, \sigma)$ at several scales, where $\sigma$ is the scale parameter. The convolution result $L(x, y, \sigma)$ can be defined as:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$$  \hspace{1cm} (1)

Where * means the convolution operation at coordinates $(x, y)$, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2) / 2\sigma^2}$$  \hspace{1cm} (2)

In order to detect stable keypoints in scale space efficiently, the method proposed by Lowe [10] is used, which uses the difference-of-Gaussian function convolved with the image. The difference of two nearby scales separated by a constant scale factor $c$ can be computed as:

$$D(x, y, \sigma) = (G(x, y, c\sigma) - G(x, y, \sigma)) * I(x, y)$$

$$= L(x, y, c\sigma) - L(x, y, \sigma)$$  \hspace{1cm} (3)

In any case for scale space feature description, the smoothed images $L$ at every scale need to be computed. Therefore, in this method $D$ can be computed by simple image subtraction.

In order to detect the extrema reliably, there is an important issue on how to determine the frequency of sampling in scale and spatial domains. Here we take the settings made by Lowe [10], 3 scales per octave and the standard deviation $\sigma$ of Gaussian $G$ is set to be 0.5.

By taking apart all sampling issues and by applying several thresholds to eliminate unreliable features, the SIFT method computes scale-space extrema $(x_i, y_i, \sigma_i)$ of the spatial Laplacian $L(x, y, \sigma)$ and samples for each of these extrema a square image patch centered at $(x_i, y_i)$, which has dominant gradients over its neighbors. Because the resulting image patches at scale $\sigma_i$ are searched basing on gradient direction, it is invariant to illumination changes. Furthermore, only local histograms of the direction of the gradient are kept, the SIFT descriptor is invariant to translation and rotation.

### 2.2 Affine-SIFT (ASIFT)

The idea of combining simulating all zooms out of the query image and normalizing rotation and translation is the main ingredient of SIFT method. Based on this idea, ASIFT method simulates the two camera axis parameters, the longitude angle and the latitude angle (which is equivalent to tilt), and then applies SIFT method to simulate scale (zoom out) and to normalize translation and rotation.

Similar to SIFT method, the sampling frequency needs to be considered because simulating the whole affine space is not prohibitive and impractical. Furthermore, a two-resolution scheme for comparing the similarity between two images needs to be used to reduce the ASIFT complexity.

#### 2.2.1 ASIFT algorithm

Step1: Simulates all possible affine distortions of the query image, where the distortions are caused by the change of camera optical axis orientation from a frontal view. The degree of distortion is dependent on two parameters, the longitude angle $\phi$ and the latitude angle $\theta$. For longitude angle $\phi$, the query image undergoes rotations. For latitude angle $\theta$, the query image undergoes subsamples with parameter $t = \frac{1}{\cos \theta}$, which means the convolution by a Gaussian with standard deviation $k\sqrt{t^2 - 1}$. The constant value $k = 0.8$ is settled by Lowe [10].

Step2: Since the efficiency of computation needed to be taken into account, the sampling steps are performed on a finite number of latitude angles and longitude angles.

Step3: All simulated images from the query image are compared by a similarity matching method (SIFT).

#### 2.2.2 Acceleration with a two-resolution scheme for ASIFT

The two-resolution scheme is used to accelerate the process of computing the similarity between two images. The main idea of this scheme is firstly selecting the affine transforms that yields well matches at low-resolution. Then it simulates images from the query and the searched images both at these selected affine transforms and at original-resolution. Finally, computes the similarity between these simulated images. The steps of two-resolution scheme are summarized as follows:

Step1: Computes the low-resolution images of the query image $u$ and the searched image $v$ by using a Gaussian Filter and a downsampling operator. The resulting low-resolution images can be defined as:

$$u' = P_F G_F u \quad \text{and} \quad v' = P_F G_F v$$  \hspace{1cm} (4)

where $u'$ and $v'$ are the low-resolution images of $u$ and $v$, respectively. $G_F$ and $P_F$ are the Gaussian Filter and downsampling operator, respectively. And the subindex $F$ represents the size factor of operator.

Step2: Applies ASIFT method to $u'$ and $v'$. 
Step 3: Selects $M$ affine transforms that yielding well matches between $u'$ and $v'$.

Step 4: Applies ASIFT method on $u$ and $v$ at the $M$ affine transforms selected by step 3. And chooses the best match among these $M$ transforms as the similarity between $u$ and $v$.

Fig. 1 shows some face examples to compare the matching ability between SIFT and ASIFT method, in which left two columns show the matching results for SIFT method and right two columns show the matching results for ASIFT method respectively. From Fig. 1 we can see that when the variation of facial pose and angle of the same person is higher, SIFT method could not find any match. And among all examples, the matching ability of ASIFT method is obviously superior to SIFT method.

2.3 The proposed method

The face image representation we adopted in this paper is ASIFT Visual Words. In this representation, visual vocabulary must firstly be generated. Visual vocabulary is generated by using hierarchical $K$-means method to cluster a large number of ASIFT descriptors. The reason we adopted hierarchical $K$-means method here is by considering both the clustering time and the clustering efficiency. When the clustering process is convergent, the centers of all clusters are forming the visual vocabulary. Now the ASIFT descriptors of every face image can be converted to a histogram form by using visual vocabulary. Suppose there are $z$ centers (say code words) $\{C_1, C_2, \ldots, C_z\}$ in visual vocabulary and there are $r$ ASIFT descriptors $\{A_1, A_2, \ldots, A_r\}$ in a face image. For each descriptor $A_j$, $1 \leq j \leq r$, we calculate the Euclidian Distances between $A_j$ and all the centers $C_i$, $1 \leq i \leq z$. Chooses the center which has minimum distance and records the index of this center in $R(A_j), 1 \leq j \leq r$. The visual words representation of this image is defined as:

$$H(i) = \frac{1}{r} \sum_{j=1}^{r} E'(j), 1 \leq i \leq z,$$

where $E'(j)$ is as follows:

$$E'(j) = \begin{cases} 1, & \text{if } R(A_j) = i. \\ 0, & \text{otherwise.} \end{cases}$$

And $H(i)$ is a histogram of length $z$ and it is also the visual words representation of this face. The distance between two visual words representations from two faces can be evaluated by Bhattacharyya distance [14].

2.4 Performance Evaluation of Face Recognition In Video Sequences

There are many schemes in face classification of video sequences, such as probabilistic majority voting and Bayes maximum a posterior scheme proposed in [15]. In both schemes, the similarity between a test image and a video sequences is computed by considering the similarities between this test image and all images in this video sequences. This is not appropriate since two face images of a person from two different facial poses may introduce lower similarity. And this will lower the overall similarity between a test image and a video sequences from the same person. In this paper, we define the similarity between a test image $w$ and a video sequences $S$ as:

$$\text{Sim}(w, S) = \max_{i} \text{Sim}(w, s_i), s_i \in S.$$  (7)

Where $s_i$ is a face image in the Video Sequence $S$. In this definition, among all the similarities between a test image and all face images in a video sequences, only the maximal similarity is used.

III. EXPERIMENTAL RESULTS

In this section, we use the popular UCSD/Honda video database and VidTIMIT video database to evaluate the performance of face recognition. In UCSD/Honda video database, there are 59 video sequences of 20 different people. There are about 300-600 frames in each video, including large pose and expression variations with significantly complex out-of-plane (3-D) head rotations. These 59 video sequences are further divided into a training subset of 20 videos and a testing subset of 39 videos. Among all members in this database, only one person did not take any testing video.
VidTIMIT video database contains 43 different people and each person has 13 video sequences. There are about 250-500 frames in each video sequences. The first 3 sequences of each person were taken under 4 regular head movements (left, right, up and down) and the last 10 sequences were taken under speaking short sentences. In our experiment, we use the first sequence as training video and use the second and the third sequences as testing videos.

For all video sequences, Viola-Jones face detector [16] was firstly applied to detect face in each frame. Then we would manually delete frames which the detected position of face is incorrect. Fig. 2 and Fig. 3 show some correctly detected face samples on these two corpora. All detected faces are preprocessed by illumination compensation [17]. In our experiment, the first 25 frames and the first 100 frames respectively from every subject’s training and testing face sequences are used for performance evaluation. And the numbers of visual phrases we used in UCSD/Honda database and VidTIMIT database are 9000 and 16384, respectively.

The proposed ASIFT Visual Words method is compared against other four classical approaches, they are LBP[18], MBLBP[19], Local Gabor Binary Pattern[20], SIFT Visual Words. The recognition rates on UCSD/Honda video database and on VidTIMIT video database are summarized in Tables I and II, respectively. From the Table I we can see that the proposed method is superior to other classical approaches.

From Table II, the recognition rate of our proposed method is also superior to other classical approaches, but the improvement of performance is not obvious. The reason is probably that the head movements are regular in this database and the facial poses are similar in training and testing videos. Furthermore, there is no facial expression change in the first 3 video sequences used in our experiment.

### IV. CONCLUSIONS

In this paper, we proposed a face recognition method which uses ASIFT visual words as image representation. We also proposed a face recognition scheme in video sequences to further improve face recognition performance.

### REFERENCES

[7] Kazuhiro Fukui, Bjorn Stenger, and Osamu Yamaguchi, “A framework for 3D object recognition using the kernel constrained...
mutual subspace method,” Asian Conference on Computer Vision (ACCV), 2006.


