Visual Language Model for Face Clustering in Consumer Photos

Wei-Ta Chu
Department of CSIE
National Chung Cheng University, Taiwan
wtchu@cs.ccu.edu.tw

Ya-Lin Lee
Department of CSIE
National Chung Cheng University, Taiwan
lylin96m@cs.ccu.edu.tw

Jen-Yu Yu
Info. and Comm. Research Labs
Industrial Technology Research Inst. Taiwan
KevinYu@itri.org.tw

ABSTRACT
For consumer photos, this work clusters faces with large variations in lighting, pose, and expression. After matching face images by local feature points, we transform matching situations into a novel representation called visual sentences. Then, visual language models are constructed to describe the dependency of image patches on faces. With the probabilistic framework, we develop a clustering algorithm to group the same individual's face images into the same cluster. An interesting observation about evaluating face clustering performance is proposed, and we demonstrate the superiority of the proposed visual language model approach.

Categories and Subject Descriptors
I.5.2 [**Pattern Recognition**]: Design Methodology – *feature evaluation and selection, pattern analysis.* I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding. I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement – *feature representation.*

General Terms
Algorithms, Experimentation.

Keywords
Face clustering, visual language model, agglomerative clustering.

1. INTRODUCTION
Recently, people take large amounts of photos to capture travel or daily life experience, with the low-cost digital cameras or mobile devices equipped with cameras. The increasing number of photos rapidly incurs great challenges in media management, retrieval, and browsing. For media management, in addition to annotate where the photos were taken and what objects were in these photos, human beings have special interests on annotating who were in photos. Face annotation is, therefore, a fundamental issue for consumer photo management. This trend can be confirmed by the development of web-based albums embedded with face annotation functions [1].

The special challenges of face recognition or clustering in consumer photos are at least twofold. First, drastic pose and lighting variations make the conventional eigenface approach fail. Although techniques of face appearance model or face alignment have been proposed for years, promising results haven't reported for consumer photos. Second, some studies were conducted to annotate faces based on not only eigenface similarity, but also some context information such as clothes [2][3]. However, accurately finding clothes regions under uncontrolled capture environments is rather an open issue.

In this paper, we propose a brave new idea that exploits visual language models to describe face similarity, and accordingly conduct face clustering by an agglomerative clustering approach. This idea is motivated by that we often say two similar persons have similar eyes, noses, mouths, etc. For example, we would describe two brothers who both have thick eyebrows, almond eyes, raised mouth, etc. For human beings, a sequence of similar parts on faces drives the perception of similarity. In this work, we develop a model that "visually" describes the similarity between two faces, based on a visual sentence expressing the face matching situation. With the help of visual language models, we cluster similar faces in an agglomerative manner, given a set of unconstrained consumer photos.

The rest of this paper is as follows. Section 2 reviews studies related to face clustering and visual language models. Section 3 describes face matching based on local feature points. In Section 4, face matching situations are transformed into visual sentence, and visual language models are constructed to conduct clustering. Experimental results are provided in Section 5, and Section 6 gives the concluding remarks.

2. RELATED WORKS
To specially tackle with face clustering for consumer photos, not only standard face recognition techniques but also external context information were utilized by previous researches. Zhang et al. [2] extracted three features from the upper part of body, face, and eyes, and proposed a Bayesian framework to describe and predict the identification of each face. Zhao et al. [3] proposed a graphical model to integrate face and clothes information. Further post-processing was developed to eliminate identification errors. In our previous work [4], we developed a module based on local feature points matching to enhance clustering performance. In addition to consumer photos, challenges derived from moving faces in videos increasingly draw attention. Tao and Pan [5] segment videos into sequences with face images in similar poses. They then identify faces for each pose-constrained sequence.

MM'09, October 19–24, 2009, Beijing, China.

Language models describe statistical characteristics of words, and have been widely studied in natural language processing. Wu et al. [6] first applied the idea to image classification, and proposed the construction of visual language models. Image patches and the extracted visual features are quantized into virtual words. Based on local feature points, the idea of visual words was proposed to present images and was used in image search [7]. Motivated by both visual language models [6] and visual words [7], we propose a new representation for describing face matching situations, and evaluate face similarity by visual language models.

3. FACE MATCHING

In contrast to eigenface approaches, we evaluate face similarity from a totally different perspective, i.e. local feature points matching. We exploit difference of Gaussian (DoG) feature detectors to locate feature points, and then describe each feature point by a SIFT (Scale-invariant feature transform) descriptor [8]. The SIFT descriptor is used because it is invariant to scale and rotation, and is robust to some degree of illumination and viewpoint changes. These are important characteristics to analyze faces in consumer photos.

Figure 1(a) shows two examples of SIFT-based matching between the same persons' face images. We can obviously see that many matches can be found on some prominent parts, such as eyebrows, eyes, nose, and mouth. These parts are the important features by which human beings recognize people. In addition, we found that these parts can be matched even illumination changes (left) or pose changes (right). By contrast, Figure 1(b) shows that the number of matches between different persons' face images is much smaller.

For the same persons' face images, feature matches often co-occur on different parts. On the other hand, although different persons' face images may have matches, the matching situations are relatively more random than that of the same persons (Figure 1(b)). It is said that a face may have similar nose and mouth with some others, but two faces likely present the same person if they both have similar eyebrows, eyes, nose, nose, mouth, and so on (Figure 1(a)). According to these observations, we develop a systematic method to present face matching situations.

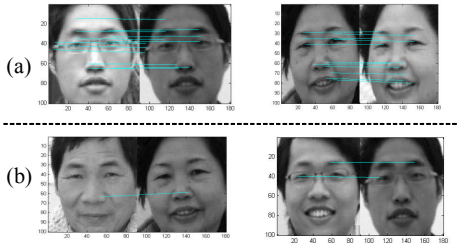

Figure 1. SIFT-based matches between (a) the same persons' face images and (b) different persons' face images.

4. VISUAL LANGUAGE MODELING

4.1 Matching Situation Representation

We first divide each face image into three regions by a ratio of 3:2:3 from up to down, i.e., the upper region, the middle region, and the bottom region, which approximately present the eyebrows and eyes, nose, and mouth. Then, a pair of faces are matched based on SIFT descriptors [8]. The matched feature points in each region are aggregated as follows.

$$s^r(i) = \sum_{k=1}^{N} s_k(i), \quad (1)$$

where $s_k(i)$ denotes the value of the i-th dimension of the k-th matched feature point in the r-th region, assuming that there are N matched feature points in this region. The value $s^r(i)$ is then normalized to guarantee that the maximum value is limited no more than 1.

For a pair of faces, we finally obtain three aggregated feature points s^1, s^2, and s^3. Based on the AT&T face database [11], we collect aggregated feature points from 400 face images, and apply the k-means algorithm to cluster similar features into groups, where each group represents a visual word [7]. The set of visual words is denoted by V in the following descriptions. Conceptually, each visual word represents distinct features in the upper, the middle, or the bottom regions of a face.

Given a pair of test faces f_p and f_q, we perform SIFT-based matching and construct aggregated feature points, quantize them into visual words, and transform the matching situation into a visual word sequence, i.e., a visual sentence, by traversing visual words from top to bottom. Let's denote the visual sentence by $S_{p,q} = \langle v_1 v_2 v_3 \rangle$, where v_i is the visual word corresponding to the aggregated feature point s^i. Figure 2 shows that two faces of the same individuals are matched, and the sequence of matched visual words can be conceptually viewed as (1) corner of the glasses; (2) tip of nose; and (3) corner of mouth. That is, corresponding to the notation above, the visual word v_1 conceptually corresponds to corner of the glasses, v_2 conceptually corresponds to top of nose, and so on.

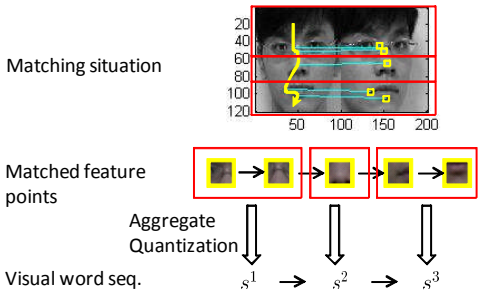

Figure 2. Matching situation in terms of visual words.

4.2 Model Training

To efficiently and effectively construct a model that describes the dependency between visual words, we make the following assumptions in visual language modeling. First, each visual word in the same visual sentence is correlated. Moreover, the dependency between visual words is generated from top to bottom, as shown in Figure 2. Therefore, the visual word proximity of a matching situation is measured by a probability form:

$$p(v_k | v_1 v_2 ... v_N) = p(v_k | v_1 v_2 ... v_{k-1}). \quad (2)$$

To further simplify this conditional probability, techniques of conventional language model suggest that each visual word only depends on its immediate neighbors. This introduces variations of model settings, depending on how many neighbors are considered. In this work, three visual language models are experimented, i.e., unigram, bigram, and trigram models. In unigram model, each visual word is considered independently. In bigram model, each visual word depends on its previous closest visual word, In trigram model, the dependency of the visual word and its previous

two closest visual words is modeled. These three models can be respectively described as follows.

$$p(v_k|v_1v_2...v_N) = p(v_k), \quad (3)$$

$$p(v_k|v_1v_2...v_N) = p(v_k|v_{k-1}), \quad (4)$$

$$p(v_k|v_1v_2...v_N) = p(v_k|v_{k-2}v_{k-1}). \quad (5)$$

To characterize different face matching situations, we construct two visual language models. The first visual language model describes the matching situations between faces of the same individuals, in a form of conditional probability distributions. The second visual language model describes the matching situation between two distinct individuals.

- Unigram

The unigram model is constructed as

$$p(v_k|C_i) = \frac{\text{Count}(v_k|C_i)}{\sum_{v\in V}\text{Count}(v|C_i)}, i = 1, 2. \quad (6)$$

The symbols C_1 and C_2 denote matching situations between faces of the same individuals and different individuals, respectively. The function $Count(v_k|C_i)$ denotes the count of the visual word v_k appearing in the visual sentences collected from C_i. In language modeling, zero probability would harm the succeeding classification process. Therefore, we assign a small prior probability for this case. Handling the zero probability problem is an age-old but important issue in natural language processing. Detailed smooth methods please refer to [9]. In this work, we utilize the toolkit provided by [9] to implement language models.

- Bigram

The bigram model describes the probability of a visual word that conditionally depends on its previous closest neighbor:

$$p(v_k|v_{k-1}, C_i) = \frac{\text{Count}(v_{k-1}v_k|C_i)}{\text{Count}(v_{k-1}|C_i)}, i = 1, 2. \quad (7)$$

Smooth methods for bigram may be more complicated than that for the unigram model. However, detailed implementation is beyond the scope of this paper, and readers are referred to [9].

- Trigram

Similarly, the trigram model is constructed as follows

$$p(v_k|v_{k-2}v_{k-1}, C_i) = \frac{\text{Count}(v_{k-2}v_{k-1}v_k|C_i)}{\text{Count}(v_{k-2}v_{k-1}|C_i)}, i = 1, 2. \quad (8)$$

The probability of a visual word that conditionally depends on its previous two closest neighbors is described.

4.3 Face Clustering

With the visual language models, we first pick up outliers and then cluster the remaining face images. Lastly, the outliers are assigned to appropriate face clusters by a specially designed method. Details of these processes are described as follows.

- Outlier Selection

Some face images may be captured in significantly varied poses, or may be too blurred due to motion or bad lighting. This kind of face image never resembles any else. Such images are viewed as outliers, and we should conduct special process for them. For a pair of faces f_i and f_j, the face likelihood ratio is defined as

$$r_{i,j} = \frac{p(S_{i,j}|M_1)}{p(S_{i,j}|M_2)}, \quad (9)$$

where $S_{i,j}$ is the visual sentence representing the matching situation between f_i and f_j, M_1 is the visual language describing matching situations between the same individual's faces, and M_2 describes matching situations between different individuals' faces.

A face image is selected as an outlier if its likelihood ratios to all other face images are below a threshold:

$$r_{i,j} < \delta, \quad \forall j, j \neq i. \quad (10)$$

- Clustering

For the face images other than outliers, they are clustered by an agglomerative process. Each face image first forms a face cluster, i.e., $F_1 = \{f_1\}$, $F_2 = \{f_2\}$, …, $F_N = \{f_N\}$. Two face clusters F_{i^*} and F_{j^*} are merged if

$$(i^*, j^*) = \arg\min_{1\le i,j\le N, i\neq j} \mathcal{H}(F_i, F_j), \quad (11)$$

$$\mathcal{H}(F_i, F_j) = \max(h(F_i, F_j), h(F_j, F_i)), \quad (12)$$

$$h(F_i, F_j) = \frac{1}{|F_i|}\sum_{f_p^{(i)}} \min_{f_q^{(j)}}(1 - p(S_{p,q}|M_1)), \quad (13)$$

where $\mathcal{H}(F_i, F_j)$ is a modified Hausdorff distance between the clusters F_i and F_j. The distance $h(F_i, F_j)$ between the face $f_p^{(i)}$ in F_i and the face $f_q^{(j)}$ in F_j is evaluated counter to the probability of being the same individual. The value $|F_i|$ denotes the number of face images in F_i.

The clustering process proceeds until the desired number of clusters have been reached.

- Outlier Assignment

For each face image in the outlier set $O = \{f_1, f_2, ..., f_M\}$, we assign each of them to one of the existing face clusters. The outlier f_i is assigned to the cluster F_{j^*} if

$$j^* = \arg\max_{1\le j\le M} \frac{1}{|F_j|}\sum_{k=1}^{|F_j|} r_{i,k}, \quad (14)$$

According to this equation, an outlier is assigned to a face cluster by checking its average likelihood ratio to existing clusters and finding the one that causes a maximum value.

5. EXPERIMENTS

We first verify how different language models affect clustering performance. Four hundred face images from the AT&T face database [11] are selected to train a unigram, a bigram, and a trigram model, respectively. For testing, we evaluate these models by clustering five sets of face images. Table 1 shows that these faces are with different scales of lighting variations, expression variations, and pose variations. The fourth and the fifth datasets are captured by amateurs in family tours. Figure 3 shows the clustering accuracy. The accuracy value of each face cluster is calculated by dividing the number of correctly clustered faces by the total number of faces in this cluster. For a dataset, the average accuracy is calculated by averaging the accuracy values of all face clusters in it, and is illustrated in Figure 3.

Figure 3 shows that we generally have the worst performance for the fourth and the fifth datasets, and the first three datasets that are captured in control environments have satisfactory performance. These results confirm that consumer photos with large variations in lighting, expression and pose harm the clustering performance. The bigram and trigram models behaves much better than the unigram model. The average accuracy values over these five datasets for unigram, bigram, and trigram models are 0.58, 0.74, and 0.73, respectively. Therefore, we adopt the bigram model for the following experiments.

In performance evaluation, we collect consumer photos recording travel or daily life, and also use a subset of photos from an open photo collection [10]. There are 17 datasets containing totally 1409 face images. The number of persons in a dataset range from two to seven. To verify the proposed method, we first investigate

the characteristics of one of the closest commercial applications – the name tags function in Google Picasa [1]. Generally, Picasa achieves high face clustering accuracy. However, the price of high accuracy is that Picasa often "over-clusters" the given photo sets. There are averagely 3.72 face clusters in the 17 datasets, but Picasa averagely segments them into 21.2 face clusters! The extreme case is that we put each face in a cluster, then we obtain 100% clustering accuracy for each cluster. Therefore, according to this observation, we argue that a good face clustering system should not only achieve high clustering accuracy, but also limit over-clustering when high accuracy is achieved.

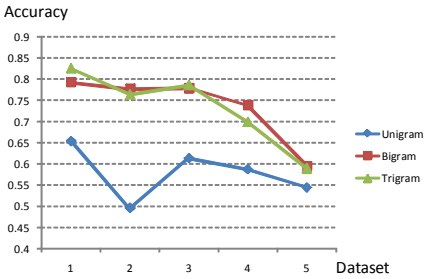

Figure 3. Comparison of different language model settings.

Table 1. Information of the test data and their characteristics.

Test dataset		# face images	# clusters	Description
1	AT&T	400	40	sLV, sEV, sPV
2	Lab faces	368	10	sLV, sEV, sPV
3	Lab daily	89	7	sLV, lEV, lPV
4	A family	42	5	lLV, sEV, lPV
5	B family	56	4	lLV, lEV, lPV

LV: lighting variation; EV: expression variation; PV: pose variation; s: slight, l: large.

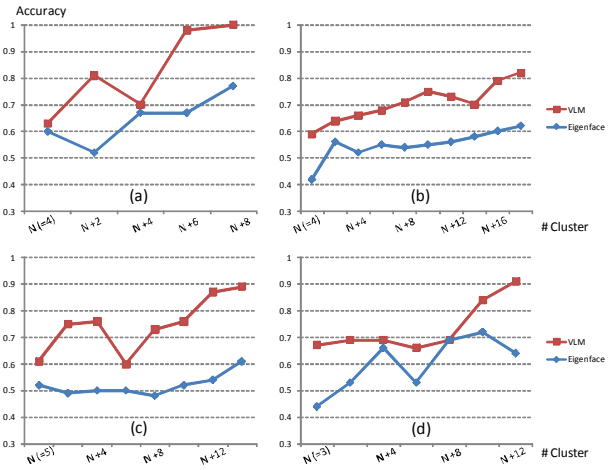

Figure 4. Cluster accuracy vs. number of face clusters in four different datasets.

Figure 4 shows the clustering performance versus the number of face clusters in our method and the conventional eigenface approach. The eigenface-based face clustering method is implemented by clustering the eigenface coefficients corresponding to each face image into a desired number of clusters. From Figure 4, we see in both methods that accuracy increases as the number of clusters grow. This trend confirms the strategy of Picasa. Moreover, our method clearly performs better than the eigenface approach. For example, in Figure 4(a), we achieve 100% accuracy when 12 clusters are allowed in a dataset that actually has 4 (N=4) different individuals. With the same setting, the eigenface approach only achieves 78% accuracy.

To quantitatively measure the clustering performance, we calculate a ratio by considering the number of face clusters when a specific clustering accuracy is achieved: $R = |F_{eig}|/|F_{VLM}|$, where $|F_{eig}|$ and $|F_{VLM}|$ are the numbers of face clusters obtained by the eigenface approach and our method that first time achieve at least 80% face clustering accuracy. After evaluating the 17 datasets, we finally get the average ratio $\bar{R}$=1.58, which means that the conventional eigenface approach over-clusters 1.58 times than the proposed visual language model approach.

6. CONCLUSION

A new viewpoint is proposed to effectively address face clustering for consumer photos, in which faces have large variations in poses, lighting, and expression. We elaborately transform matching situations between faces into visual sentence representation, and construct visual language models to describe the dependency of different parts of faces. Based on the probabilistic framework, an agglomerative clustering approach is used to group the same individual's faces into the same cluster. The experimental results demonstrate superior performance and confirm the trend of developing a practical face clustering application.

7. ACKNOWLEDGEMENT

This work was partially supported by the National Science Council of the Republic of China under grants NSC 97-2221-E-194-050.

8. REFERENCES

[1] Picasa web albums, http://picasaweb.google.com

[2] Zhang, L., Chen, L., Li, M., and Zhang, H.J. 2003. Automated annotation of human faces in family albums. In Proc. of ACM Multimedia, pp. 355-358.

[3] Zhao, M., Teo, Y.W., Liu, S., Chua, T.-S., and Ramesh, J. 2006. Automatic person annotation of family photo album. In Proc. of CIVR, pp. 163-172.

[4] Anonymous review

[5] Tao, J. and Tan, Y.-P. 2008. Efficient clustering of face sequences with applications to character-based movie browsing. In Proc. of ICIP, pp. 1708-1711.

[6] Wu, L., Li, M., Li, Z., Ma, W.-Y., and Yu, N. 2007. Visual language modeling for image classification. In Proc. of MIR, pp. 115-124.

[7] Sivic, J. and Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In Proc. of ICCV, 2, pp. 1470-1477.

[8] Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, vol. 60, no. 2, 91-110.

[9] Clarkson, P.R. and Rosenfeld, R. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In Proc. of ESCA Eurospeech.

[10] The Gallagher Collection Person Dataset, http://amp.ece.cmu.edu/people/andy/GallagherDataset.html

[11] AT&T Laboratories Cambridge, http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html