An Experimental Study on Content-Based Face Annotation of Photos

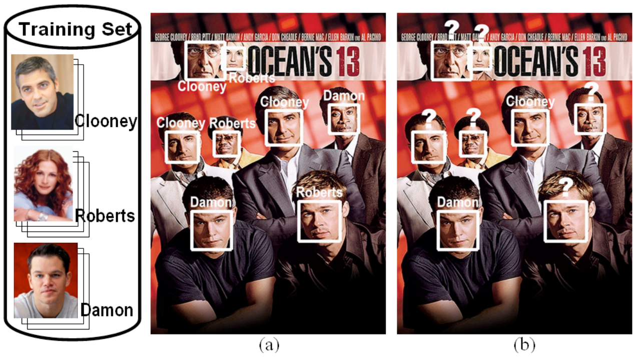
Mei-Chen Yeh, Sheng Zhang, and Kwang-Ting Cheng

Abstract— **Face annotation of photos, a key enabling technology for many exciting new applications, has been gaining broad interest. The task is different from the general face recognition problem because the dataset is not constrained—an unlabelled face may not have any corresponding match in the training set. Moreover, faces in real-life photos have a significantly wider variation range than those in the conventional face datasets. We designed and conducted a thorough experimental study to understand the efficacy of face recognition methods for annotating faces in real-world scenarios. The findings of this study should provide information for various design choices for a practical and high-accuracy face annotation system.**

I. Introduction

In recent years, the number of consumer photos has increased rapidly with the wider availability of devices with cameras and the proliferation of free image storage web sites. People can readily share and browse pictures with negligible cost and effort. To organize the expanding volume of personal photo collections, automatic face annotation [1, 3, 7, 8, 9, 14, 17, 19] can be used to label faces with names, which enables efficient photo searching. More recently, online face annotation systems have been commercially launched, such as Picasa Web Albums [27] and Riya [28].

Face annotation is generally considered a combined problem of face detection and face recognition. Given an input image, a face detector is first applied to locate face areas and each localized face is assigned a name based on the most similar identity in the training set. However, this formulation—the input of face recognition is the outcome from face detection—creates new, specific challenges for the traditional face recognition problem. First, names associated with a few training images initially provided by users are relatively limited. Thus, it's necessary to mark some faces as *unknown*. These unknown faces might come from people whom we don't know (e.g. pedestrians or strangers in pictures taken during vacation) or from known people whose image we have not yet labeled to train the system. Second, the localized face from the face detector could be a false positive (i.e. not a human face). Traditional face recognition methods always find the most similar face in the gallery set for every query face, which might result in an incorrect annotation result, as shown in Fig. 1a.



(a) (b)

Fig. 1. Example of face annotation. Left: the training set consists of three people. Middle: each detected face is assigned a name based on the training set. Right: preferred annotation.

The goals of face annotation are two-fold: to achieve great accuracy in recognizing people of interest and to reject any unknowns as well as false positives. However, the task of recognizing faces among known people alone, especially in candid photos, remains highly challenging in spite of the vast amount of research conducted on this subject [25]. Numerous factors other than identity—such as lighting conditions, facial expressions, poses, and partial occlusions—, change the way a face appears in an image. Rejecting faces that do not appear in the training set is even more challenging. Unknown faces share features with known faces since both are images of faces. Moreover, they lack coherent patterns, thus making them difficult to characterize. Most existing face recognition methods learn discriminative information for differentiating among people who appear in the training set. Their ability to reject unknowns has not been thoroughly evaluated and remains an open question.

In order to understand the efficacy of existing, substantially sophisticated methods for annotating faces in real-world scenarios and further identify the best combination of technologies for designing a practical annotation system, we conducted an experimental study that consists of three components. First, we conducted a comparative evaluation with combinations of several state-of-the-art face representation and recognition methods. We identified the use of local features combined with the Linear Discriminant Analysis (LDA)-based recognition method achieves promising recognition performance—93.83% accuracy on the traditional face recognition dataset FERET [16] and 71.69% accuracy on the more challenging photo dataset LFW [8]. Second, we proposed a new measure for evaluating a method's ability to reject unknowns. With this measure, we study the tradeoff between the rejection rate on unknown

M. Yeh (myeh@csie.ntnu.edu.tw) is with the Computer Science and Information Engineering Department, National Taiwan Normal University, Taipei, Taiwan. S. Zhang (s_zhang@psych.ucsb.edu) is with the Psychology Department, University of California, Santa Barbara, CA 93106 USA. K. -T. Cheng (timcheng@ece.ucsb.edu) is with the Electrical and Computer Engineering Department, University of California, Santa Barbara, CA 93106 USA.

faces and the recognition rate drop on known faces, which should be carefully considered when designing a practical annotation system. Last but not least, we built a face dataset collected from family photo albums. It is more challenging than other real-world imagery because family members tend to look alike and these sets of photographs tend to consist of people from the same racial background. With this dataset, we are able to conduct cross-set evaluations by combing it with other life photo sets such as the news photos in [3, 8]. To the best of our knowledge, we are the first to examine methods' abilities to reject unknown and non-face images.

Note that the use of multiple modalities is effective for face annotation and there have been extensive studies that address this problem [3, 7, 9, 14, 17]. However, one of our target applications—labeling faces with names in consumer photos—does not necessarily contain additional text cues. As a result, we focused on methods that solely rely on visual information. It is important to understand the level of performance that can be achieved solely by using visual cues, but the performance can be improved whenever information from other modalities is available.

In the rest of the paper, we start by describing the face representations and face recognition algorithms evaluated in this study. In Section 3 we present a method for measuring both recognition and rejection performances, followed by the description of datasets used in the evaluation. Finally, we demonstrate the comparative evaluations in Section 5 and conclude the paper.

II. Recognition Methods

Content-based face recognition generally consists of three steps—pre-processing, feature extraction and classification. In this section we briefly discuss methods examined in this study.

A. Pre-processing

The measured performance of face recognition methods can be significantly biased if non-facial areas (e.g. background, hair) are used [18]. Features extracted from those non-facial areas are not completely uncorrelated to the face, and may thus provide accidental, additional hints for recognition. For example, Shamir showed that the recognition accuracy could achieve 99% on the YaleB face dataset solely by using features from a small background area [18]. To minimize the use of such non-facial features, we extracted features from the tightly bounded face area, as shown in Fig. 2b.

Face representation is, in general, categorized into two classes—global and local [25, 26]. Global features, such as pixel intensities, are extracted from the whole face area. Local features, on the other hand, are concatenated features extracted from several local facial parts. Fig. 2 shows our flow for deriving local regions. First, we employed a face detector [21] to obtain the exact face rectangle. Following this step we used Active Appearance Models [13] to find coordinates of landmarks[1] and determined five local regions on the face. The face area is then normalized and aligned based on those coordinates.

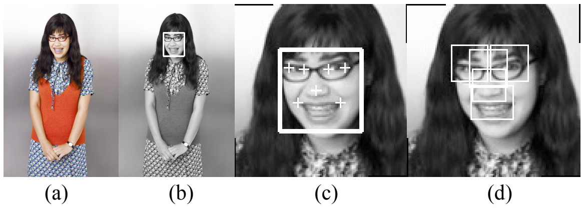


Fig. 2. Flow of extracting local regions: (a) input image; (b) detected face area; (c) landmark coordinates; (d) local regions.

B. Face Description

Face description is as critical as face modeling for a successful recognition system. Extraction of face features must be automatic and efficient. Moreover, feature descriptors should be compact and discriminative. In this study we examined three methods: pixel intensities [2, 20], Gabor filters [10], and Scale Invariant Feature Transform (SIFT) [12]. In the following we describe details of our implementation. To derive a fixed-length descriptor from varying sizes of face images, we rescaled the face region (or the facial regions) to a fixed resolution. For Gabor filters, we used the integral Gabor-Haar transformation [10], which is essentially using Haar-like feature extraction on raw Gabor features. We implemented 2-D Gabor filters with 5 scales, 8 orientations, and 5 Haar-like features, providing 200 face features for describing an area. SIFT features have recently been demonstrated effective for matching faces [19]. We applied SIFT on the center of each facial region, thus providing 640 features per face. Table I summarizes the feature descriptions and their dimensions.

Table I
Feature Descriptions and Their Dimensions

	Global	Local
Pixel Intensities	3600	4240
Gabor filters	200	1000 (200*5)
SIFT	N/A	640 (128*5)

C. Feature Selection and Classification

Face annotation applications require a recognition approach that can handle the small sample size problem (i.e. the data dimensionality is greater than the number of training samples) and the multi-category classification problem (i.e. the number of people of interest is greater than two) with efficient computation. It is also desirable to involve less parameter tuning for optimizing the recognition results. Moreover, it should offer the flexibility of adding additional training images, either for new or existing names. Considering those criteria, we examine template-based recognition approaches such as Principal Component

[1] Active Appearance Models do not necessarily give correct landmark coordinates. We do not use any hand-labeled ground truth data of landmarks in our experiments. All feature extraction processes are automatic.

Analysis (PCA) [20], several variants of Linear Discriminant Analysis (LDA) [5], and a Support Vector Machine (SVM) [4]. To overcome the small size problem in LDA, several methods have been proposed in the literature. In this study we implemented the Fisherface [2], Regularized LDA [6], Dual-Space LDA [22], and LDA/FKT [24] methods. The classification is accomplished by measuring the cosine similarity on the projection space. We also compare them with the Nearest Neighbor (NN) classifier to determine the amount of performance gain by applying these dimensionality reduction schemes. Table II summarizes the details of our implementations.

TABLE II
DESCRIPTIONS OF FEATURE SELECTION METHODS

Method	Description
NN	We used the cosine measure for comparing two vectors.
PCA	Also known as Eigenface [20]. We selected top eigenvectors that keep 95% of the eigen-energy.
PCA+LDA	Also known as Fisherface [2]. We first applied PCA to keep 95% eigen-energy, followed by LDA.
Regularized LDA [6]	We computed the generalized eigenvectors of $(S_w+\mu I)^{-1}S_b$ where we set $\mu = 0.05$.
DSLDA [22]	We select *C*-1 eigenvectors from the principal, and the null space of S_w, respectively, where *C* is the number of people under recognition.
LDA/FKT[24] or KFKT	We used subspace 1 with a dimensionality of *C*-1. If this subspace did not fully exist, we used the kernel extension of this method [11].
SVM	We used the implementation of LIBSVM [4] with linear kernel, and a pair-wise coupling scheme for multi-category classification.

III. MEASURE FOR REJECTING UNKNOWNS

Many classifiers yield a set of scores between an unseen pattern and class models—numeric values that represent the likelihood of the unseen pattern belonging to a class. These scores can either be strict probabilities or some general, un-calibrated proximity values or distances. For example, NN-based face recognition approaches produce such scores by using a distance function in a (transformed) feature space. A natural solution for rejecting a pattern is to set a threshold on the maximum score or on a measure that combines these scores, such as the p-value, which quantifies the confidence of the classification decision.

A face annotation system can be evaluated from the retrieval perspective if we care about how the system ranks those unlabelled faces, including known and unknown faces, given training faces from known persons. Precision and recall are two widely used measures for evaluating the quality of results [8, 9, 14, 19]. By introducing an unknown face dataset in the testing phase, we are ready to calculate the precision and recall by ranking images in both the testing set and the unknown set based on their scores to each identity model. However, these measures are sensitive to the amount of irrelevant data (the size of the unknown dataset), and may not directly relate to the primary recognition performance.

Inspired by the signal detection theory, we propose to use a receiver operating characteristics (ROC)-like measure, as shown in Fig. 3, in which the recognition rate on known faces (plotted on the Y-axis), the rejection rate on unknown faces (plotted on the X-axis), and their tradeoffs can be simultaneously displayed. The origin (0, 0) in the figure implies a classifier rejects every pattern. On the other hand, a classifier that does not have the rejection option corresponds to the point with x = 1. The measure also depicts a classifier's ability to rank the known faces (positive instances) relative to the unknown faces (negative instances).

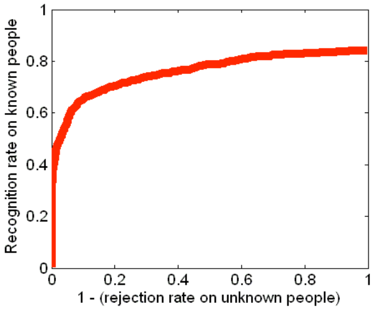


Fig. 3. The recognition-rejection curve.

This measure has a few attractive properties. First, it is insensitive to class skew (e.g. changes in the proportion of known and unknown faces) or error costs. Face annotation applications may contain a wide range of class skews, and the ratio of known and unknown faces changes as users label more and more images. It is important that the measure used is invariant to such changes. Secondly, this measure provides a performance spectrum. For situations in which negative samples dominate, the performance area with a great rejection rate becomes more interesting. Third, it is easy to reduce this 2-D depiction to a single scalar value for comparing classifiers. Common methods such as the area under the curve (AUC) and equal error rate (ERR) can be used to derive the value.

IV. DATASET

We conducted experiments on one standard database—FERET [16], and two photo datasets—news photos and consumer photos. As shown later in this paper, we observed a significant disparity between the results reported in research papers and those of real-world field tests. However, the use of FERET helps validate that our implementation achieves baseline performances.

We followed the same setup in [23] and created the FERET subset. This dataset consists of 1000 images of 200 individuals and each of them has five images. Fig. 4 shows a few sample images. It has variations in facial expression, illumination, and pose; however, we used a different method for cropping the face regions. In [23], each face region was cropped based on the labeled location of eyes and mouth, and contained hair and background. Instead, we cropped the photo using a much tighter face region (i.e., non-face areas are excluded), resulting in a more challenging dataset.

The first photo dataset, referred to as Faces in the Wild, was collected by Berg et al. [3] from Yahoo! News photos during a period of roughly two years. The original dataset has 30,281 detected faces, grouped by their clustering algorithm into 14,808 clusters with 77% accuracy. After discarding clusters with less than 20 elements and merging those clusters corresponding to the same person, 126 clusters remain. We manually removed near-duplicate images and reduced the number of clusters to 74. This resulted in a dataset of 3,523 images of 74 people, in which the number of images per person ranged from 13 to 699. Fig. 4 shows samples from this dataset. Our building process is very similar to that of the Labeled Faces in the Wild (LFW) dataset [8]. We did not use LFW because it is designed for studying the pair matching problem. However, we can use its protocol for face verification to demonstrate the performance of our best combination of methods, which is later shown in the experimental section.

The second photo dataset, named the Family dataset, was collected from four family albums[2]. A few samples are shown in Fig. 4 (c). This dataset contains 1,158 images of 18 people. Although only 18 people are in this dataset, it is a challenging task to annotate these photos because family members tend to look alike. Besides, family albums often include images of a person collected over a long period of time (e.g. children progressing from infancy to teenage; adults aging over time, etc.) and tend to consist of people from the same racial background. Table III summarizes the datasets used in our study.

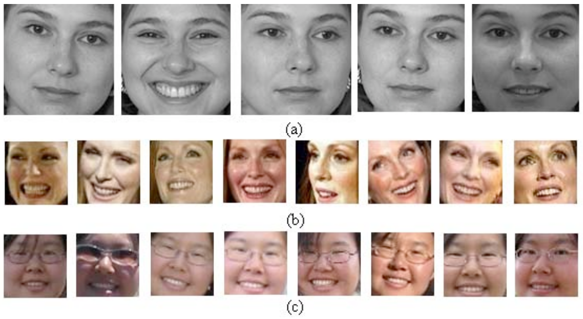


Fig. 4. Sample images in (a) FERET, (b) Yahoo! News, and (c) Family datasets.

TABLE III
STATISTICS OF FACE DATASETS

Name	# {people}	Size	Description
FERET	200	1,000	Standard dataset
Yahoo! News	74	3,523	News photos
Family Albums	18	1,158	Consumer photos

V. EXPERIMENTS

A. Setup

Practically speaking, a user typically provides only a small number of training samples initially. More and more training samples would become available over time. It is important to observe how the recognition accuracy changes with respect to the number of training samples. The accuracy curve could provide the information needed for suggesting the number of required training samples for a specific target accuracy. In our experiment, we randomly select N training images for each identity. The value of N ranges from 2 to 10 for Yahoo! News and from 2 to 30 for Family Albums. The remaining images serve as the testing set. For each selection of N, we repeat sampling 10 times, and show the mean and the standard deviation of the average recognition accuracy across categories[3]. For the FERET dataset, as suggested in [23], we used two images—frontal neutral and frontal smiling faces—for training, and the remaining three images for testing.

B. Comparative Evaluation

We first compare different combinations of face description and face modeling. In Section 2, we discussed five types of face descriptions: global/local pixels, global/local Gabor filters, SIFT, and several dimensionality reduction or classification methods: NN, PCA, LDA, and SVM. To illustrate the contribution of each component to the overall recognition performance, we also show the results for each component.

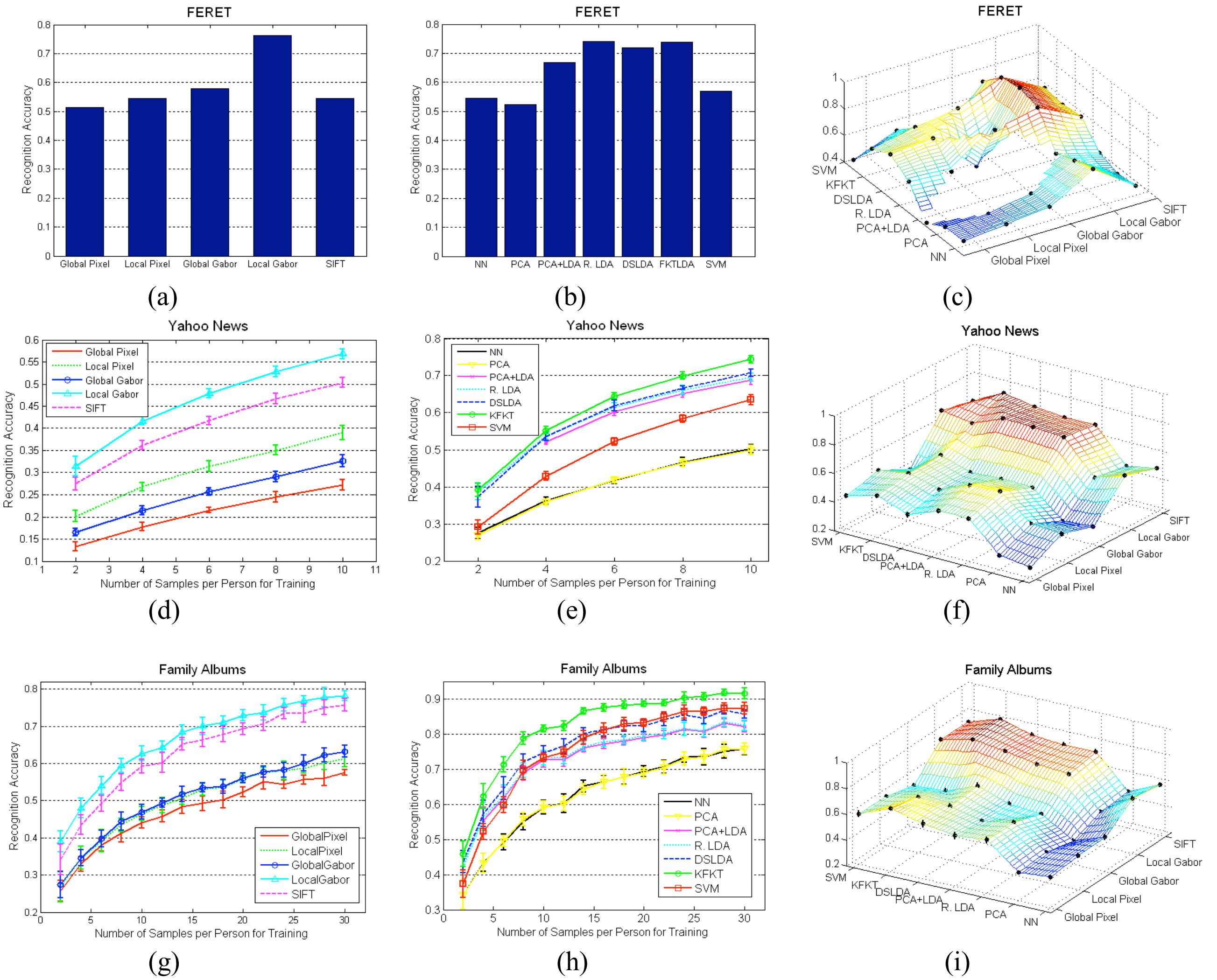
The recognition accuracy rates across datasets are shown in Fig. 5. Rows are results for the dataset of FERET, Yahoo! News, and Family Albums. The first column illustrates the comparison of descriptions. We applied the NN classifier to report the recognition accuracy rates. The comparison of feature selection methods is shown in the second column. In these experiments we used the SIFT features. The last column demonstrates the accuracy rates using all 35 combinations of descriptions and methods evaluated in this study given 10 training images per person. In Fig. 5c, the hole in this graph is caused by the fact that accuracy rates using PCA+LDA with certain descriptions are unavailable since the feature dimension in the transformed space after PCA is less than that required by LDA.

We first observe a significant disparity between FERET and the photo datasets**.** Given two training images per person, the combination of local Gabor filters and DSLDA can recognize 200 people in FERET with 93.83% accuracy; however, the recognition rates drop to 40.89% and 46.39% when the other datasets are used. This result suggests that the performance reported using standard face datasets tends to be too optimistic for our target applications. We observe a dramatic performance drop (half the accuracy rate) when real-life photos are used.

Local features outperform global features in all datasets, and among these local features, local Gabor and SIFT features perform the best. Considering Fig. 5(a), (d) and (g), those features extracted from automatically detected facial parts perform better than those extracted from the whole face—regardless of which descriptions are used. The large

[2] http://lbmedia.ece.ucsb.edu/research/annotation/annotate.htm

[3] Since each category contains different number of testing images, the overall performance may be biased to those categories with more images without normalizing accuracies across categories.



(a) (b) (c)

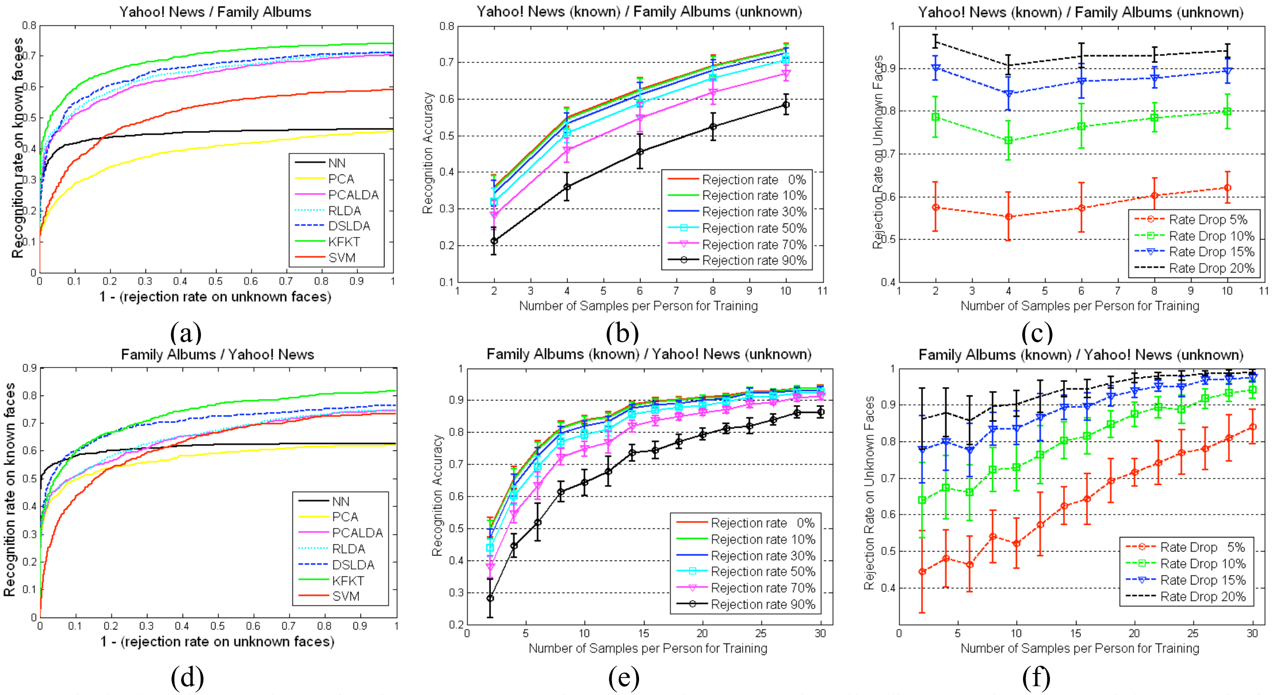
(d) (e) (f)

(g) (h) (i)

Fig. 5. Comparative evaluation of descriptions and methods on three datasets: FERET (a)-(c), Yahoo! News (d)-(f), and Family Albums (g)-(i). The first column shows comparison between descriptions. The second column shows comparison between feature selection methods. The last column shows the accuracy of all combinations.

performance gain (18.5% in FERET, 14.96% to 24.12% in Yahoo! News, 11.57% to17.47% in Family Albums) is apparent especially when using the Gabor features. Among local features, local Gabor features seem to be the best; however, *SIFT's ability is obscured in the FERET experiments*. The performance gain of local Gabor features versus SIFT is reduced from 21.83% for the FERET dataset to less than 5% for the photo datasets. As we will show, the use of SIFT features and KFKT gives the best performance for both the Yahoo! News and Family Albums sets. This result suggests that SIFT is a good feature for describing faces in photos. It might also reflect the fact that SIFT was originally designed for reliable matching between different views of 3-D objects—thus, it is more robust to the variations common in photo collections.

One interesting question rises following the discussion of local features: which facial part is most discriminative? Psychophysical studies indicate that human recognition identifies the eyes as being the best feature for recognizing faces, followed by the mouth, and then the nose [15]. However, previous work on machine vision ranks these parts differently. In [26], the authors argue that machines favor facial parts that contain the least noise. Thus, the nose is probably the best feature for distinguishing a person since it is less noisy than eyes and mouth as the shapes of noses are less likely to be distorted. Our study suggests a different result—the discriminative power of facial parts for machines depends upon the way we describe them. Table IV shows the recognition accuracy rates for FERET using each facial part alone with the KFKT method. The numbers shown in red indicate the ranking. We observed similar results regardless of which classifier is applied, and the same observation was made when the photo datasets were used. As shown in [26], Gabor features are discriminative for describing *stable* parts such as nose and the region between the eyes; however, *noisy* parts are more discriminative when SIFT is used (compare among accuracy rates in the bottom column). This result suggests we may describe different parts by applying different descriptions to them.



(a) (b) (c)

(d) (e) (f)

Fig. 6. Evaluation including an unknown face dataset: (a)-(c) use Yahoo! News as known set and Family Albums as unknown set, and vice versa (d)-(f). The first column shows the recognition-rejection curves. The second column shows recognition rates with various rejection rates. The last column shows the rejection rates with different recognition rate drop.

Fig. 5 (b), (e) and (h) show a feature selection method does improve the separation of features from different identities. The performance gain ranges from 11.7% to 24.14% depending on the number of training samples. In all experiments, LDA-based approaches achieve the best performance, followed by SVM, PCA and NN. It is not surprising that those discriminative learning approaches in which class labels are used in the training phase perform better than unsupervised approaches. Moreover, the LDA/FKT produces the best results since it truly maximizes the Fisher criterion [5], and the kernel extension of this method (KFKT) ensures the existence of such discriminative subspace where the Fisher criterion is equal to infinity. Fig. 5 (h) shows encouraging results on the use of KFKT and SIFT—81.57% and 90.43% recognition rates, given 10 and 24 training samples per person.

The recognition accuracy rates using all combinations with 10 training samples per person are shown in Fig. 5 (c), (f) and (i). For the FERET dataset, the best recognition rate, which is 93.83%, is achieved by using local Gabor features and DSLDA. This result is better than the best result reported in [23], which was 81.2%. Experiments on two photo datasets consistently suggest that the best result is achieved by using a combination of SIFT and KFKT. The accuracy rate is 74.42% for Yahoo! News dataset and 81.57% for the Family Albums dataset, both with a small accuracy variation (1.01% and 0.98%). This combination also achieves great performance (near 60% improvement) using the face verification protocol and the LFW dataset. The estimated mean accuracy is 71.69%±2.18%, while the baseline performance reported in [8] is 12.7% ±0.45%.

As shown previously in this paper, local Gabor and SIFT features deliver comparable overall performance, but are inconsistent in ranking the different facial parts. If we concatenate both features, this fused descriptor boosts the accuracy rate to 76.86% for the Yahoo! News photos and to 85.83% for the Family Albums dataset.

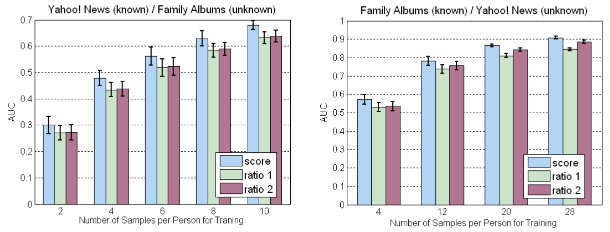
TABLE IV
PERFORMANCE OF FACIAL PARTS

	Left eye	Right eye	Bet. eye	Nose	Mouth	All
Gabor	73.0% 4	74.8% 3	78.8% 2	82.8% 1	70.5% 5	91.2%
SIFT	68.8% 2	69.3% 1	49.3% 3	37.3% 4	31.2% 5	73.8%

C. Cross-set Evaluation for Rejecting Unknowns

In this section, we investigate a method's ability to reject unknown faces by using the recognition-rejection measure introduced in Section 3. In this experiment we require an unknown dataset in the testing phase to calculate the rejection rate. Since both the Yahoo! News and the Family Albums are constructed from candid photos and the people appearing in these two dataset are completely different, we conducted a cross-set evaluation that used one group as the known dataset for training the system, and the other as the unknown dataset for testing the rejection ability.

The evaluation results are shown in Fig. 6. Each row shows three different charts using the same known and unknown face sets. Fig. 6 (a) and (d) demonstrate the recognition-rejection curves in one experiment given 10 training samples per person. Again, we used SIFT for feature description. The figures show how the recognition rate varies



(a) (b)

Fig. 7. The area under the curve (AUC) using different scores.

with different rejection rates for the unknowns. It is clear that the recognition rates increase while the rejection rates decrease, and are bounded by the case in which no rejections are allowed. Most papers in the literature report the end points of these curves, which provide no information on the false positive rate. In other words, we don't know how a classifier performs when samples outside the training set are present. In fact, in the early stage of training, unlabelled faces are usually dominated by large numbers of negative instances. Thus, the performance in the far left-hand side of the curve becomes more interesting.

In our target application, we need to compare the recognition rates at both zero and non-zero rejection rates. A good method's recognition rate at a higher rejection rate should be as close as possible to that at the zero rejection rate. For example, in Fig. 6d the PCA+LDA, Regularized LDA, DSLDA, and SVM methods achieve similar recognition performance at the zero rejection rates. However, DSLDA is considered better than the other methods as, at a higher rejection rate, it achieves a higher recognition rate than others. Among these methods NN has the best shaped curve. That is, it declares positive classification only when strong evidence is present and makes few false positive errors. Indeed, most feature selection methods learn a discriminative space solely by looking at a strictly constrained training set, and, thus, they might not be general enough to differentiate the known and unknown faces. This result suggests the need to combine both the discrimination and the generalization ability—for example, the use of a training set as well as an unknown face dataset for discriminative learning might help enhance the generalization ability.

We now select the KFKT method and show more performance details using its recognition-rejection curve. Fig. 6 (b) and (e) demonstrate the average recognition accuracy rates with different rejection rates across a range of training samples. In both sets, the drop in the recognition rates is fairly small when the rejection rate is below 50%. Fig. 6 (c) and (f) show the rejection rates with the recognition rate drop ranging from 5% to 20%. However, the recognition rate drops significantly when a larger rejection rate (>50%) is required.

The recognition-rejection measure can also be used to evaluate those methods that combine scores. It reflects the separability of score distributions between known and unknown faces. In the previously mentioned experiments we set a threshold on the highest score between an unseen pattern and the class models. An alternative method is to set the threshold on the ratio of scores (e.g. the ratio of the highest score to the second highest score). If there are multiple training images for one class, we must make sure the second-highest score is for a different class from the first. As explained in [12], using a score ratio makes more sense since correct matches need to have the score significantly greater than that of a "closest" incorrect match. Fig. 7 shows the AUC for three different scoring methods using the SIFT+KFKT method. The blue bar represents the AUC that directly uses the score for thresholding, while the other two use the score ratios, s1/s2 and 1-(s2/s1), where s1 and s2 are the highest and the second highest scores. These results indicate that for SIFT+KFKT, thresholding based only on the score would be a good choice as it achieves a greater AUC.

VI. Conclusion

In this paper we study the face annotation problem for real-life photos. It is different from the general face recognition problem because faces from photos contain a wider range of variations in appearance, and an unlabelled face may not have any matching face in the training set. We implemented, experimented, and analyzed the performance of several state-of-the-art face description and modeling methods. A few remarks can be drawn from this study: 1) the use of real-life datasets is important. We observe a significant disparity between performance using standard face datasets and photo datasets. Moreover, a method's ability can be obscured if only standard datasets were used for evaluation; 2) both face description and face modeling methods are crucial to achieving good recognition performance. Our study suggests combining SIFT features and KFKT classification could provide a good baseline method for real-world testing; 3) local facial parts have different discriminative power rankings depending upon the description used; and 4) a figure-of-merit is introduced which should be more suitable for studying the face annotation problem. We hope the experimental results derived from this study are beneficial for the design of a practical and high-accuracy face annotation system.

References

[1] O. Arandjelovic and A. Zisserman, "Automatic face recognition for film character retrieval in feature-length films," *CVPR*, vol. 1, pp. 860-867, 2005.

[2] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *PAMI*, 19(7), pp. 711-720, 1997.

[3] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth, "Who's in the picture?" *NIPS*, 2004.

[4] C.-C Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, 2nd Edition*. John Wiley and Sons, 2000.

[6] J. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, 84:165-175, 1989.

[7] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Automatic face naming with caption-based supervision," *CVPR*, pp. 1-8, 2008.

[8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: a database for studying face recognition in

unconstrained environments," University of Massachusetts, Amherst, Technical Report 07-49, 2007.

[9] V. Jain, E. Learned-Miller, and A. McCallum, "People-LDA: Anchoring Topics to People using Face Recognition," *ICCV*, pp. 1-8, 2007.

[10] J. Li, T. Wang, and Y. Zhang, "Face Recognition using Feature of Integral Gabor-Haar Transformation," *ICIP*, pp. 505-508, 2007.

[11] Y. Li and M. Savvides, "Kernel Fukunaga-Koontz transform subspaces for enhanced face recognition," *CVPR*, 2007.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 60(2), pp. 91-110, 2004.

[13] I. Matthews and S. Baker, "Active appearance models revisited," *IJCV*, 60(2), pp. 135-164, 2004.

[14] D. Ozkan and P. Duygulu, "A graph based approach for naming faces in news photos," *CVPR*, pp. 1477-1482, 2006.

[15] S. E. Palmer, *Vision Science*, *Photons to Phenomenology*. Cambridge, MA: MIT Press, 1999.

[16] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *PAMI*, 22(10), pp. 1090-1104, 2000.

[17] S. Satoh and T. Kanade, "Name-It: association of face and name in video," *CVPR*, pp. 368-373, 1997.

[18] L. Shamir, "Evaluation of face datasets as tools for assessing the performance of face recognition methods," *IJCV*, 79(3), pp. 225-230, 2008.

[19] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: video shot retrieval for face sets," *CIVR*, pp. 226-236, 2005.

[20] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, 3(1), 1991.

[21] P. Viola and M. Jones, "Robust real-time face detection," *IJCV*, 57(2), pp. 137-154, 2004.

[22] X. Wang and X. Tang, "Dual-space linear discriminant analysis for face recognition," *CVPR*, pp. 564-569, 2004.

[23] J. Yang, D. Zhang, J.-Y. Yang, and B. Niu, "Globally maximizing, locally minimizing: unsupervised discriminant projection with applications to face and palm biometrics," *PAMI,* 29(4), pp. 650-664, 2007.

[24] S. Zhang and T. Sim, "Discriminant subspace analysis: a Fukunaga-Koontz approach," *PAMI*, 29(10), pp. 1732-1745, 2007.

[25] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey," *ACM Computing Surveys*, 35(4), pp.399-458, 2003.

[26] J. Zou, Q. Ji, and G. Nagy, "A comparative study of local matching approach for face recognition," *IEEE Trans. on Image Processing*, 16(10), pp. 2617-2628, 2007.

[27] http://picasaweb.google.com/

[28] http://www.riya.com/