Prototype Based Feature Learning for Face Image Set Classification

Mingbo Ma, Ming Shao, Xu Zhao and Yun Fu
Electrical and Computer Engineering
Northeastern University, Boston, MA, USA
{mingboma, zhaoxu, yunfu}@ece.neu.edu, mingshao@ccs.neu.edu

***Abstract*— Recognizing human face from image set has recently seen its prosperity because of its effectiveness in dealing with variations in illumination, expressions, or poses. In this paper, inspired by the prototype notion originating from cognition field, we obtain discriminative feature representation for face recognition by implementing prototype formation on image set. The contribution of this paper is twofold: first, we propose to use prototype image sets as a common reference to sufficiently represent any image set with the same type; in addition, we propose a novel framework to extract image set's features through hyperplane supervised by max-margin criterion between any image set and prototype image set. The final features are summarized through pooling technique along the prototype image sets. We experimentally prove the effectiveness of the method through extensive experiments on several databases, and show that it is superior to the state-of-the-art methods in terms of both time complexity and recognition accuracy.**

I. Introduction

Face recognition has been a widely studied research topic in computer vision field for over two decades [28]. A working face recognition system should be capable of effectively dealing with the variations of illumination, pose, occlusion, expression and so forth. Conventional face recognition is usually performed on the basis of single query images, which are limited in covering rich variations of face appearance under complex environment. Recently, image set based classification has been introduced to computer vision applications especially for face recognition [8], [4], [16], [1], [23], [11], [12], [15]. Face recognition from image sets utilizes one or more image sets, where every single query is collectively represented by a set of diverse images.

Compared with the single image based recognition system, image set classification can extract more discriminative and comprehensive information by exploring the temporal relationship between the images from consecutive video frames [8]. However, using image set for classification also brings some challenges because it may manifest the changes of view-point, illumination and deformation. And, the model has to be carefully designed to exploit the semantic relationship between individual images. For example, in [8], [9], a sparse approximated nearest neighbor is proposed to measure distance between image sets. In this model, an image within a set is sparsely related with other samples. In [12], relationships between set members are explored with linear discriminant analysis. Canonical correlation is used to measure distance between different image sets.

To avoid explicitly exploring semantic relationship between set members as presented in [12], [8], [9], in this paper, we propose a novel representation for image set based face recognition. The idea is inspired by the *Prototype Formation* originally proposed in psychology and cognition field [17], [18]. Research in cognition reveals that human being categorize the objects based on hierarchical prototypes. Every class of observable objects and abstractive concepts have their prototypes. These prototypes help people to recognize and differentiate the world, and therefore prototype formation is a critical skill for category learning. Especially for human face recognition, related psychological experiments have verified that there exist prototypes for face recognition, which are gradually changed with the evolution of the external environment. Fig. 1 gives an intuitive explanation about the proposed prototype based image set representation.

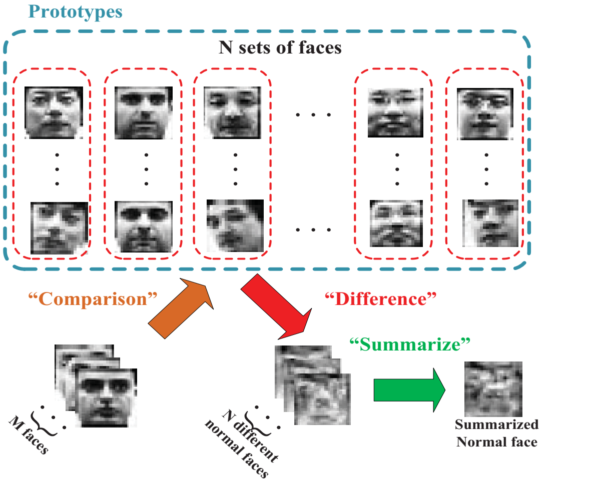

Fig. 1. Intuition explanation of our prototype inspired representation. Each face image set make comparisons with the pool of prototypes. Each prototype is composed of a set of face image. The differences are recorded and summarized for the following classification.

In the proposed framework, image set classification works with the instantiation of the concept of prototype. According to the prototype theory, a generic face image prototype pool is required because it is a common point of reference to quantitively measure the differences between image sets. Abiding by this baseline, one should be able to differentiate subtle disparity between any pair of face image sets.

Motivated by such considerations, we embed the prototype formation into image set based face recognition. In Fig. 2, the framework of the proposed approach is illustrated. In this framework, multiple generic face image sets are firstly built to act as the prototypes. Both probe (query) and gallery image sets are aligned against the prototypes to measure the

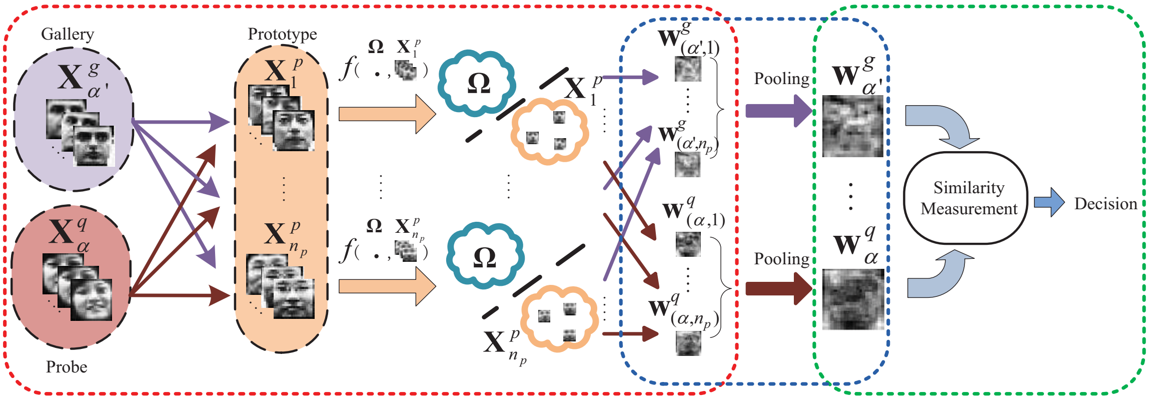

Fig. 2. Framework of the proposed prototype inspired image set classification for face recognition. Ω represents the sample space including both gallery and probe image sets and f is the classifier learning function. See section III for details.

differences with the prototype set, and these differences are utilized later to formulate features for both probe and gallery image sets. In our method, the differences are described by the hyperplanes between the prototypes and image set, because the hyperplane reside in the same high dimensional space with the original image and contain rich discriminative information. It can be obtained by maximizing the margin between both kind of sets. Through a pooling operation, a collection of such hyperplanes gives rise to an informative representation of this set of face images.

In sum, the contributions of this work are as follows. First we propose a prototype based approach to face image set recognition. It's novel to motivate this kind of application under the framework of prototype learning. Secondly, to implement this idea, we design a feature learning strategy by exploring the discriminative information contained in the hyperplanes between prototypes and different face sets. No matter what the semantic or temporal relationships between samples are, the information has already been automatically incorporated in the representation. At the same time, weight within the hyperplane space indicates the capability of differentiation. Finally, the experiments conducted in Honda/UCSD and Hong Kong PolyU NIR data sets demonstrate the superior performance of our proposed approach.

II. Related Work

The concept of prototype has been defined in Eleanor Rosch's study "Natural Categories" [17]. It was first defined as a stimulus associated with a category and then redefined as the most central member of a category. It directly resulted in set-theoretic approaches of "extensional or intensional semantics". Prototype theory has been applied in linguistics, as part of the mapping from phonological structure to semantics [5], [10]. Intuitively, prototype more likely abstracts out the central tendency from the experienced examples, and then use it as a basis for categorization decisions.

Prototype effect in face recognition has been demonstrated in cognition research [18], [21], [3]. Psychological experiments reveal that prototype plays critical roles in recognizing human face because people are prone to recognizing the face corresponding to the central value of a series of observed faces. It means that prototype could provide a measure and baseline to recognize unseen face. It's worth mentioning that a computational approach is proposed for face recognition by measuring prototype similarities [13], where each face image is described as a vector of kernel similarities to a collection of prototype face images. Inspired by the same prototype theory but with a completely different computational methodology with [13], we represent a set of face images by measuring the hyperplane between face images set and prototype sets. In addition, it should be notified that some similar concepts, i.e., associate set [25], has been proposed to approach the impact factors such as pose and illumination, where probe and gallery sets are approximately aligned by a comprehensive intermediate face data set. However, their feature extraction methods are still under the conventional framework.

When working with image set classification, the main concern is how to extract set information and then effectively represent it for classification. In [8], an image set is represented as an affine hull associated with the number of image samples and their mean. Although not explicitly using a prototype model, in this representation, the affine hull model is used to implicitly construct a prototype to account for unseen face images. Similar to [8], [9] , in [4], each image set is characterized by a convex geometric region spanned by its feature points, and set dissimilarity is measured by geometric distances between convex models. In [12], image sets are transformed by a discriminant function and then compared by canonical correlations. In their work, representation and similarity measure between sets are integrated into a framework of discriminant-analysis of canonical correlations. It should be noted that manifold and subspace learning give rise to a typical class of approaches to represent and measure image sets [6], [11], [1], [12], [23].

Different from the image set representation mentioned above, in this work, we construct informative representation of a face image set by searching hyperplanes to maximize the margins between prototypes and image sets. In this way, the newly designed features for image sets naturally benefit the classification task because of the implicit discriminative measurement inside.

III. Methdology

In this section, we introduce the proposed methodology. The whole framework is illustrated in Fig. 2 where the first part enclosed in the red rectangle shows the pipeline of extracting discriminative information from input image sets against prototypes. The second part enclosed by blue rectangle is the formation of final representation of each image set by pooling operation. Classification results can be obtained by employing nearest neighbor, as illustrated in the green rectangle enclosed part.

A. Max-Margin Feature Learning From Prototypes

Without loss of generality, in this paper, we represent a face image as a vector $\mathbf{x} \in R^D$ by concatenating the intensity of all pixels column by column. Thus an image set with n face images can be denoted as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in R^{D \times n}$. In the following, we mainly work with three types of face image sets, namely, the prototype set $\mathbf{X}^p$, the probe set $\mathbf{X}^q$ and the gallery set $\mathbf{X}^g$.

In this work, we aim at learning a discriminative feature representation of an image set for face recognition. Inspired by the prototype theory, we explore the representation of image sets by viewing generic face image sets as the prototypes, which are represented as $\mathcal{X}^P = \{\mathbf{X}_1^p, \mathbf{X}_2^p, \cdots, \mathbf{X}_{n_p}^p\}$, where n_p is the number of prototypes, and the superscript p indicates that this is the prototype set. It needs to be noted that the size of each image set may be different. To achieve informative and discriminative representation, each image set in both probe sets $\mathcal{X}^Q = \{\mathbf{X}_1^q, \mathbf{X}_2^q, \cdots, \mathbf{X}_{n_q}^q\}$ and gallery sets $\mathcal{X}^G = \{\mathbf{X}_1^g, \mathbf{X}_2^g, \cdots, \mathbf{X}_{n_g}^g\}$ are described by measuring the differences with the prototype set $\mathcal{X}^P$, where n_q and n_g are the sizes of probe sets and gallery sets respectively. Suppose I_{proto} denotes the set of identities in prototype image sets, and therefore I_{probe} and I_{gallery}, the probe and gallery image sets. To learn the difference of people's faces more efficiently, the identities of people from prototypes are totally different from the identities in both gallery and probe image sets. Mathematically, we have $I_{\text{gallery}} \cap I_{\text{proto}} = \emptyset$ and $I_{\text{probe}} \cap I_{\text{proto}} = \emptyset$, but $I_{\text{gallery}} \cap I_{\text{probe}} \neq \emptyset$.

Intuitively, a unique hyperplane that best separates an input set and a prototype set is applicable for discriminative features of the input since all distinct parts of the input set are now fully represented by the prototype set. This process is clearly shown in Fig. 2, where we try to find the hyperplane between each image set in $\mathcal{X}^Q$ and $\mathcal{X}^G$, and all of the prototypes in $\mathcal{X}^P$. In principle, any classifier, linear or non-linear, could be used to form the classifier boundary. Specifically, for a probe set $\mathbf{X}_\alpha^q$, and a prototype $\mathbf{X}_\beta^p$, where $\alpha \in I_{\text{probe}}, \beta \in I_{\text{proto}}$ the linear classifier boundary is:

$$\mathbf{w}^q_{(\alpha,\beta)}\mathbf{x} + b = 0, \tag{1}$$

where $\mathbf{x}$ is a vector in the face vector space spanned by $\mathbf{X}_\alpha^q$ and $\mathbf{X}_\beta^p$, and $\mathbf{w}^q_{(\alpha,\beta)}$ is the normal of the classifier boundary, which can be visualized as a "normal face"(Fig. 3) for a discriminative representation. Therefore, it could be reasonably used for part of features of $\mathbf{X}_\alpha^q$.

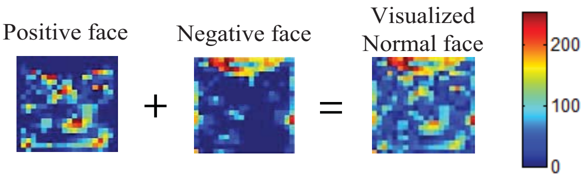

Fig. 3. Normal face has two components: positive face and negative face. Brighter color corresponds higher pixel value

Here, we consider the popular linear SVM that maximizes the margin between each pair of image set, and its non-linear version can be easily approached through kernelization. Following the canonical SVM formulation, given probe and prototype image sets the hyperplane between these two can be derived through maximizing margin in-between by optimizing the following objective function [19]:

$$\min_{\mathbf{w}^q_{(\alpha,\beta)},\xi,b} \frac{1}{2}\left\|\mathbf{w}^q_{(\alpha,\beta)}\right\|_2^2 + C\sum_{i=1}^{n_p+n_q}\xi_i, \tag{2}$$

where C controls the relative trade-offs between constraint violation and margin maximization, and ξ_i is a slack variable. The optimization is subject to the constrain:

$$\begin{aligned} y_i\left(\mathbf{w}^q_{(\alpha,\beta)} \cdot \mathbf{x}_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0, \\ i = 1, \cdots, n_p + n_q, \end{aligned} \tag{3}$$

where y_i are the labels. Once upon the hyperplane, which is represented by its normal vector $\mathbf{w}^q_{(\alpha,\beta)}$, is determined, we can view it as a informative feature representation of the input image set $\mathbf{X}_\alpha^q$ against the prototype $\mathbf{X}_\beta^p$.

Essentially, the normal vector illustrates how the corresponding hyperplane balance the separation between two sets and constrains violation by optimizing the angle of hyperplane in D-dimensional space. For each dimension, SVM supervises the processing of selecting a weight for this particular dimension to separate the two classes with minimum cost. Since our input face images are well aligned, each pixel will roughly corresponding to certain area of face. Therefore, all the pixels in the images are attached with certain sematic meaning. Different weights on different pixels from $\mathbf{w}^q_{(\alpha,\beta)}$ illustrate the importance of those corresponding pixels. The magnitude of weights imply the discriminant ability of the pixels, the higher the magnitude, the better the separation ability. When several local pixels from the same region are given high magnitudes, we are notified that this part is of great importance in differentiating two sets. As it is shown in Fig.2, $\mathbf{w}^g_{(\alpha',1)}$ and $\mathbf{w}^q_{(\alpha,n_p)}$ showcase the difference of $(\mathbf{X}^g_{\alpha'}, \mathbf{X}^p_1)$ and $(\mathbf{X}^q_\alpha, \mathbf{X}^p_{n_p})$. We found that $\mathbf{X}^g_{\alpha'}$ is the image set from a people with deep eyes, and both $\mathbf{X}^p_1$ and $\mathbf{X}^p_{n_p}$ from Asian without deep eyes. The normal face indeed highlights these significant differences, which will be emphasized more by pooling operation.

Without loss of generality, we use $\mathbf{w}$ to denote a general normal face, for either probe set or gallery set. Apparently, as a real value vector, the normal face $\mathbf{w}$ consists of positive and negative values, both of which are of great importance in describing the classifier boundary. To explicitly show the impacts of both positive and negative values in an image, we take the following operations on $\mathbf{w}$:

$$\mathbf{w}^+ = \begin{cases} \mathbf{w} & \text{if } \mathbf{w} > 0 \\ 0 & \text{if } \mathbf{w} \leq 0 \end{cases}, \mathbf{w}^- = \begin{cases} -\mathbf{w} & \text{if } \mathbf{w} < 0 \\ 0 & \text{if } \mathbf{w} \geq 0 \end{cases} \quad (4)$$

and name $\mathbf{w}^+$ and $\mathbf{w}^-$ as positive and negative normal face, respectively. In Fig.3, we visualize a normal face from the gallery set. As it is shown in Fig.3, we find that positive face $\mathbf{w}^+$ shows the difference of nose, eyes and lips, and the negative face $\mathbf{w}^-$ mainly illustrates the difference of forehead and boundary. To incorporate two parts of differences together, we add positive and negative together, and form so called visualized normal face $\hat{\mathbf{w}} = \mathbf{w}^+ + \mathbf{w}^-$. Note that this is actually different from the "normal face" $\mathbf{w}$ we use as the feature of the image set since normal face contains both positive and negative values.

B. Pooling Derived Image Set Representation

Given different prototypes, the comparison in the last subsection will yield different results. For example, one subject from the probe set has a pair of small eyes, a big nose and a thick lip. Comparing his face with the second subject's face in the prototype set, who has a pair of big eyes, a big nose and a thick lip, we find that the most discriminative feature between these two people is the size of their eyes. Once again, when we compare the first subject in the probe set with the third one in the prototype set who has a pair of small eyes, a small nose and a thick lip, however, the conclusion is different with the previous: The size of the noise becomes the most discriminative feature.

From [7], we know human have several familiar faces in memory, in this paper called prototypes. When people compare the first subject's face with more different faces in memory, more conclusions will be drawn, and their combination, called pooling operation in this paper, will form a unique descriptor for the first subject. It should be aware that the features extracted by normal face could be sparse since all the human faces share many common features but unique ones are few. Therefore, instead of taking every feature into account, SVM will eliminate these common features in the comparison with prototypes. This property will reduce the classifier's confusion when we calculate the between-distance/similarities of face images in the recognition step.

Being aware of these, the framework in this subsection aims at formulating a compact face image set representation by summarizing diverse characteristics yield in comparing with prototypes. For a given probe set $\mathbf{X}^q_\alpha$, against all the prototypes, we achieve a set of hyperplanes $\{\mathbf{w}^q_{(\alpha,1)}, \cdots, \mathbf{w}^q_{(\alpha,n_p)}\}$, where n_p is the size of the prototype sets. To further refine the extracted over-complete features, we conduct pooling operation on the obtained hyperplanes:

$$\mathbf{w}^q_\alpha = \frac{1}{n_p}\left(\sum_{i=1}^{n_p}\left(\mathbf{w}^q_{(\alpha,i)}\right)^\eta\right)^{\frac{1}{\eta}}, \quad (5)$$

where η is a factor indicating different pooling strategies. For example, when $\eta = 1$ it corresponds to the average pooling, and when $\eta \to \infty$ it corresponds to the max pooling. Eq. 5 aggregates the activations of $\mathbf{w}_{(\alpha,i)}$ to obtain an D dimensional vector $\mathbf{w}_\alpha$ as the final representation of probe image set $\mathbf{X}^q_\alpha$. Essentially, Eq. 5 maps a set of real values to a single real value and retain the most significant part. In image classification, max and average are two common pooling strategies [2] to learn the compact features and max pooling works especially well with sparse image coding [22], [24], [26]. In this work, we take advantage of the average pooling due to its robustness against several variations, e.g., pose, illumination, expression. Detailed discussion can be found in the experiment section.

C. Recognition

For a probe face image set $\mathbf{X}^q_\alpha$, the task of recognition is to acquire its identity. After we obtain the prototype based representation for all the gallery image sets $\mathcal{X}^G = \{\mathbf{X}^g_1, \mathbf{X}^g_2, \cdots, \mathbf{X}^g_{n_g}\}$, gallery images are also mapped to a discriminative feature space based on prototypes $\{\mathbf{w}^g_1, \mathbf{w}^g_2, \cdots, \mathbf{w}^g_{n_g}\}$. As the green rectangle enclosed part shown in Fig. 2, the recognition is straightforward. One can use any similarity metric to measure the distances between the probe sets and gallery sets, e.g., Euclidean distance and any state-of-the-art multi-class classifier could be applied here for direct classification, e.g., nearest-neighbor. Then final classification objective function can be written as:

$$I_\alpha = \arg\min_i \|\mathbf{w}^q_\alpha - \mathbf{w}^g_i\|_2 \quad i \in I_{\text{gallery}}. \quad (6)$$

IV. Experimental Evaluation

This section showcases comparisons of our algorithm with SANPs [9], AHISD [4] and CHISD [4], three state-of-the-art methods on image set based classification. We first divide the whole data set into three non-overlap subsets, namely, prototype image, gallery image, and probe image sets. To make a fair comparison with other methods, we use the same gallery (training) and probe (testing) images in all methods, and prototype sets are not included in either probe or gallery images. Therefore, they do not directly account for comparison results. Then, the proposed algorithm is applied on two different data sets to generate discriminative features for both probe and gallery sets.

Note that in all experiments, faces are aligned based on centers of two eyes, and histogram equalization is imposed to each raw image $\mathbf{x}$ before it is vectorized and stored as a column in data matrix $\mathbf{X}$. The hyperplane is solved through SVM implementation used in [24], and people may use libSVM to further save the running time. The penalty term C in Eq. 2 is set to 1. The frame lengths are equal for both probe and gallery sets. In the following part, we conduct experiments on two different data sets, i.e., Honda/UCSD data set and Hong Kong PolyU NIR face data set.

A. Honda/UCSD Data Set

There are 59 video sequences of 20 different subjects in Honda/UCSD data set [14]. Different poses and expressions appear across different sequences of each subject. Each

TABLE I

IDENTIFICATION RATES ON HONDA/UCSD DATASET (IN %).

Method/ Length	20		60		100		Full length	
	accuracy	time cost	accuracy	time cost	accuracy	time cost	accuracy	time cost
SANPs [9]	80	15.52s	91.43	29.75s	**94.29**	50.2s	**100**	210.27s
AHISD [4]	82.85	**0.811s**	85.71	**4.2s**	88.57	**10.6s**	85.71	42.73
CHISD [4]	82.85	16.67s	85.71	80.13s	91.43	140.11s	88.57	320.14s
Our method	**91.43**	14.51s	**94.29**	18.5s	**94.29**	19.16s	97.14	**24.8s**

image set corresponds to a video sequence. The faces in the video are detected by the algorithm proposed in [20] and then resized to gray-scale images of size 20×20. The lengths of the sets vary from 12 to 645. In this experiment, 9 sequences from 5 people are selected as our prototype image sets. Then 15 sequences from 15 people are randomly chosen as gallery sets which are the training samples in the compared algorithms and the rest 35 sequences of the same 15 people are set as probe sets which are the testing samples in the compared algorithms. There is no identification overlap between prototype sets and gallery or probe sets.

In real-world applications, usually the face detection algorithms lose tracking of a face and only the first a few images are available for classification. Therefore, we report results using all frames as well as a limited number of frames. More specifically, we showcase the experiments by setting an upper bound of maximum length of frames from 20 to 100. When a set contains fewer frames than the upper bound, all the images are used for classification. Table I summarizes the performance of both the compared algorithms and the proposed algorithm in this paper. It is concluded that our algorithm performs better in terms of accuracy in most of cases, and significantly saves running time when we use the full length of frames.

B. Hong Kong PolyU NIR Data Set

Hong Kong PolyU NIR data set [27] contains 335 subjects and there are 100 images from each subject, and all the samples were collected by a real time NIR face capture device. The related version of Hong Kong PolyU NIR data set we use in our paper contains 55 subjects, each of which comprises six expressions, i.e., anger, disgust, fear, happiness, sadness and surprise, and different poses. Similar to two previous experiments, we keep 5 random subjects as prototype sets, and treat the rest subjects as gallery and probe sets. The size of each image is fixed at 32×32.

Experimental results are shown in Table II, from which we can see that the accuracy of our method performs best in 3 cases and the time complexity is very close to AHISD, but significant fast when frame length increases. Again, we find that our method is not sensitive to the size of video frame, in terms of accuracy and time complexity. Recall that our method beats the rest in Honda/UCSD data set with significant advantages when the video frame size is small.

C. Other Issues

1) Selection of Prototype: Intuitively, different selection of prototype image sets return different normal faces. Fig.4(a) and Fig.4(b) show how the classifier performs with different numbers of frames and prototypes. From Fig.4(a), we notice that bigger image set size is appreciated by the classifier when the size of prototype is fixed since it covers more variations. Fig.4(b) illustrates that increasing the size of prototype assists the recognition task. The prototype can be recognized as pre-knowledge in our framework. Obviously, for a single probe, more comparisons with different prototypes generate more normal faces, and therefore high quality summarized normal face. When we compare the result in both Fig.4(a) and Fig.4(b), we find another interesting phenomena: compared with increasing the size of prototype, increasing the size of the each image set in prototype achieves better results. This might be understood in this way: comparisons with fewer prototypes could construct good enough summarized normal face containing certain characteristics, while recognition in real-world applications indeed needs more within-class variations.

2) Pooling Strategy: It is notified that in our two experiments, probe set, gallery set, or prototype could be under very different environments, and image set representation through one prototype might be disturbed by impact factors, e.g., illumination. If max pooling is applied, not only significant features will be preserved, but also large areas of shadows or specularities. On the other hand, average pooling will overcome this by considering features from all the prototypes. Only feature that are unique to all the prototypes can be retained after average pooling. To show the effect of different pooling strategies, we also conduct an recognition experiment by varying pooling factor η in Honda/UCSD data set. Two different frame lengths are used, namely 20 and 50, and all the other configurations are the same with previous experiments. Results are shown in Fig. 4(c) with η increasing, the performances of both 20 and 50 frames group descend, and after $\eta = 8$, the trend tends to be flatten. This conforms to the analysis beforehand.

V. CONCLUSION

In this paper, we proposed a novel, intuitively straightforward and computational efficient framework for image set based face recognition. First, we proposed to use comparisons with prototypes as discriminative features for image sets. These comparison results, essentially the hyperplanes gained by maximizing the margin between the image sets and prototypes, bears informative as well as semantic meaning for a single individual. Second, average pooling strategy is adopted to summarize all the comparison results and formulate what we call "summarized normal face". Then

TABLE II

IDENTIFICATION RATES ON POLYU NIR DATASET (IN %).

Method/ Length	5		10		30		40	
	accuracy	time cost	accuracy	time cost	accuracy	time cost	accuracy	time cost
SANPs [9]	82	4239s	**92**	15836s	94	42784s	**100**	47361s
AHISD [4]	88	**7.92s**	90	**10.97s**	94	17.73s	**100**	41.45s
CHISD [4]	86	12.63s	**92**	59.31s	**98**	197.93s	**100**	374.11s
Our method	**90**	10.71s	**92**	11.26s	96	**14.11s**	**100**	**19.12s**

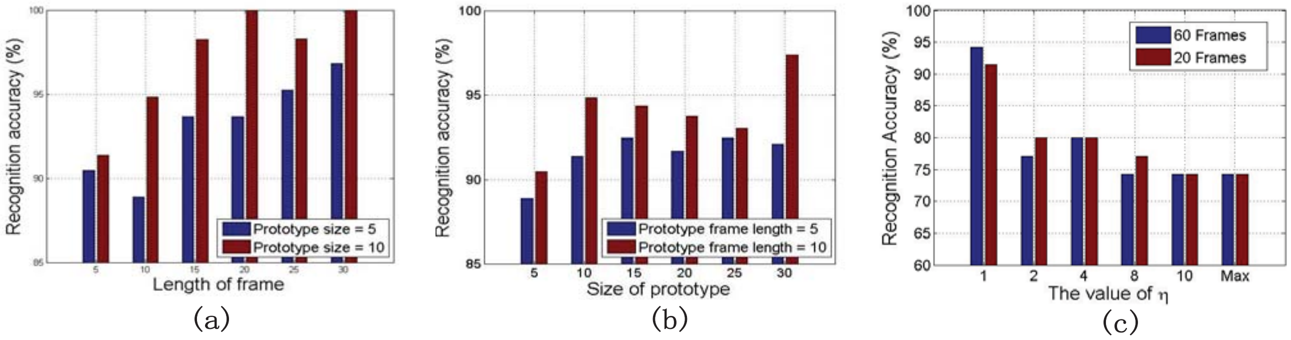

Fig. 4. Recognition performances by changing different criterions: (a) comparison between different prototype size; (b) comparison between different prototype frame length; (c) Identification rates in Honda/UCSD data set with different η

summarized normal faces are feeded to state-of-the-art multi-class classifier for final decisions. Extensive experiments on two face data sets prove the effectiveness of the proposed method on both accuracy and time complexity.

VI. ACKNOWLEDGEMENT

This research is supported in part by the NSF CNS 1135660 and 1314484, Office of Naval Research award N00014-12-1-0125, Air Force Office of Scientific Research award FA9550-12-1-0201, and IC Postdoctoral Research Fellowship award 2011-11071400006.

REFERENCES

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, volume 1, pages 581–588. IEEE, 2005.

[2] Y.-L. Boureau, N. L. Roux, F. Bach, J. Ponce, and Y. LeCun. Ask the locals: Multi-way local pooling for image recognition. In *IEEE ICCV*, pages 2651–2658, 2011.

[3] R. Cabeza, V. Bruce, T. Kato, and M. Oda. The prototype effect in face recognition: Extension and limits. *Memory & cognition*, 27(1):139–151, 1999.

[4] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573. IEEE, 2010.

[5] L. Coleman and P. Kay. Prototype semantics: The english word lie. *Language*, pages 26–44, 1981.

[6] W. Fan and D. Yeung. Locally linear models on face appearance manifolds with application to dual-subspace based classification. In *CVPR*, volume 2, pages 1384–1390. IEEE, 2006.

[7] C. Frowd, F. Skelton, C. Atherton, M. Pitchford, G. Hepton, L. Holden, A. McIntyre, and P. Hancock. Recovering faces from memory: the distracting influence of external facial features. *Journal of Experimental Psychology: Applied*, 2012.

[8] Y. Hu, A. Mian, and R. Owens. Face recognition using sparse approximated nearest points between image sets. *IEEE TPAMI*, (99):1–1, 2011.

[9] Y. Hu, A. S. Mian, and R. A. Owens. Sparse approximated nearest points for image set classification. In *CVPR*, pages 121–128, 2011.

[10] H. Kamp and B. Partee. Prototype theory and compositionality. *Cognition*, 57(2):129–191, 1995.

[11] T. Kim, J. Kittler, and R. Cipolla. Incremental learning of locally orthogonal subspaces for set-based object recognition. In *BMVC*, pages 559–568, 2006.

[12] T. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE TPAMI*, 29(6):1005–1018, 2007.

[13] B. Klare and A. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE TPAMI*, 2:6, 2011.

[14] K. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*, page 313–320, 2003.

[15] X. Li, K. Fukui, and N. Zheng. Boosting constrained mutual subspace method for robust image-set based object recognition. In *IJCAI*, pages 1132–1137. Morgan Kaufmann Publishers Inc., 2009.

[16] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara, and O. Yamaguchi. Recognizing faces of moving people by hierarchical image-set matching. In *CVPR*, pages 1–8. IEEE, 2007.

[17] E. Rosch. Natural categories. *Cognitive psychology*, 4(3):328–350, 1973.

[18] R. Solso and J. McCarthy. Prototype formation of faces: A case of pseudo-memory. *British Journal of Psychology*, 72(4):499–503, 1981.

[19] V. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York Inc, 2000.

[20] P. Viola and M. Jones. Robust real-time face detection. volume 57, pages 137–154, 2004.

[21] G. Wallis, U. Siebeck, K. Swann, V. Blanz, and H. Bülthoff. The prototype effect revisited: Evidence for an abstract feature model of face recognition. *Journal of Vision*, 8(3), 2008.

[22] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*.

[23] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, pages 429–436. IEEE, 2009.

[24] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[25] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *CVPR*, pages 497–504. IEEE, 2011.

[26] K. Yu, Y. Lin, and J. L. Erty. Learning image representations from the pixel level via hierarchical sparse coding. In *CVPR*, 2011.

[27] B. Zhang, L. Zhang, D. Zhang, and L. Shen. Directional binary code with application to polyu near-infrared face database. *Pattern Recognition Letters*, 31(14):2337–2344, 2010.

[28] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003.