# Discriminatively Learned Convex Models for Set Based Face Recognition

Hakan Cevikalp  Golara Ghorban Dordinejad
Eskisehir Osmangazi University
Machine Learning and Computer Vision Laboratory
hakan.cevikalp@gmail.com, golaradordinejad@gmail.com

## Abstract

*Majority of the image set based face recognition methods use a generatively learned model for each person that is learned independently by ignoring the other persons in the gallery set. In contrast to these methods, this paper introduces a novel method that searches for discriminative convex models that best fit to an individual's face images but at the same time are as far as possible from the images of other persons in the gallery. We learn discriminative convex models for both affine and convex hulls of image sets. During testing, distances from the query set images to these models are computed efficiently by using simple matrix multiplications, and the query set is assigned to the person in the gallery whose image set is closest to the query images. The proposed method significantly outperforms other methods using generative convex models in terms of both accuracy and testing time, and achieves the state-of-the-art results on three of the five tested datasets. Especially, the accuracy improvement is significant on the challenging PaSC, COX and ESOGU video datasets.*

## 1. Introduction

Face recognition is an important computer vision problem that has many applications in various fields. Initially, single images are used for face recognition, but more recently, set based methods have begun to dominate the field mostly because face image sets allow to model the variability of the individuals' appearances. For set based face recognition, both gallery and query sets are given in terms of sets of images rather than a single image. The classification system must return the individual whose gallery set is the most similar to the given query set. Face recognition methods using image sets are also more practical owing to the fact that they usually do not require any cooperation from the subjects. However, despite these advantages, traditional classifiers such as the support vector machines (SVMs), classification trees, $k$-nearest neighbor classifier, etc. cannot be used directly for set based recognition.

There are two important factors that determine the success of the set based face recognition methods: the models used to approximate the face image sets, and the distance metric used to measure the similarity between these models. A variety of different models were used to approximate face image sets. As a pioneering work for set based recognition, Yamaguchi et al. [34] used linear subspaces to approximate image sets, and canonical angles between subspaces are used to measure the distance between them. Another way of dealing with image set based classification is to consider each sample set as a point in a Grassmannian manifold. Hamm and Lee [10] used Grassmannian discriminant analysis on fixed dimensional linear subspaces. Wang and Shi [30] proposed kernel Grassmannian distances to compare image sets. More recently, manifolds of symmetric positive definite (SPD) matrices are used to model images sets, and the similarities between these manifolds are computed by using different Riemannian metrics such as Affine-Invariant metric or Log-Euclidean metric [20, 18, 19]. Cevikalp and Triggs [3] introduced affine and convex hulls to approximate image sets, and geometric distances between these models are used to measure the similarity. Different variants of affine and convex hulls have been proposed in [16, 36]. Among these, Sparse Approximated Nearest Points (SANP) of [16] enforces the sparsity of samples used for affine hull combination. Similarly, [36] used regularized affine hull models that require the minimization of L2-norms of affine hull combination coefficients during computing the smallest distances between image sets. Wang et al. [31] proposed a method to learn more compact affine hulls when the affine hulls of different classes overlap. More recently, new extensions [32, 37] of these methods used the so-called collaborative representations for affine and convex hull models. In contrast to the traditional methods using an independent affine and convex hull for each image set, these methods approximate all gallery sets by using a single affine or convex hull, and the query set is labeled by using the reconstruction residuals computed from only individual gallery sets. Other representative methods using sparse models in image set based recognition can be found in [9, 8, 7]. Most of the

aforementioned methods have kernelized versions that can be used to approximate nonlinear face models.

More recently, [13, 11] proposed a deep learning framework to estimate the nonlinear geometric structure of the image sets. They trained an Adaptive Deep Network Template for each image set to learn the class-specific models, and then the query set is classified based on the minimum reconstruction error computed by using those prelearnt class-specific models. Hayat et al. [12, 14] used a linear SVM classifier for approximating the distances between query and gallery sets. In a similar manner, Cevikalp and Yavuz [5] replaced the linear SVM classifier with a more suitable polyhedral conic classifier, that can return polyhedral acceptance regions for set based recognition. Lastly, there are also some related face verification and identification methods using face image sets [23, 35, 26, 22]. For example, Liu et al. [22] use multitask joint sparse representation algorithm for video-based verification. Liu et al. [23] and Rao et al. [26] use deep neural network based methods to find high quality discriminative face image frames within the image sets to improve the accuracy and speed of the face identification systems. In a similar manner, Yang et al. [35] combine a CNN network and an aggregation module to create a discriminative image set model by using high-quality image frames for video based face recognition.

**Motivation and Contributions:** With a few exceptions, majority of the set based face recognition methods are generatively learned methods which are built based on the assumption that face image sets can be represented by models created by using only the samples of those sets. Therefore, these methods focus on different models such as linear/affine subspaces, Grassmannian manifolds, SPD manifolds, etc., that best fit to the gallery sets, and they learn the model of each class by independent of other classes in the gallery. However, it is a very well-known fact that discriminative methods mostly yield to much higher accuracies compared to generative methods on classification problems. Inspired by this fact, this paper introduces a hybrid method that finds models to approximate face image sets by combining generative and discriminative methods. To this end, we approximate the face image sets by discriminative affine/convex hulls that best fit to an individual's image set, but at the same time are as far as possible from image sets belonging to other people in the gallery. As opposed to the discriminative classification methods [12, 14, 5], which require online training of a classifier by using a large-scale dataset, learning of discriminative models are implemented offline in the proposed methodology. Once the discriminative models are learned, the classification of query sets requires some simple matrix multiplications which can be accomplished very efficiently. Therefore, the proposed method is very fast compared to other discriminative methods as demonstrated in the experiments.

The remainder of the paper is organized as follows: A brief review of generatively learned affine/convex hull models are given in Section 2. We introduce the proposed method in Section 3. Section 4 summarizes experimental results. Lastly, our conclusions are given in Section 5.

## 2. Image Set Classification Based on Generatively Learned Affine/Convex Hulls

Let the face image samples be $\mathbf{x}_{ci} \in \mathbb{R}^d$, where $c = 1, \ldots, C$ indexes the $C$ image sets (individuals) and $i = 1, \ldots, n_c$ indexes the $n_c$ samples of image set $c$. [3] approximates image sets with a convex model (either an affine or convex hull) and the query image set is assigned to the class with the closest gallery set.

### 2.1. Generative Affine Hull Models

In this method, image sets are approximated by the affine hulls of their samples by ignoring face images of other classes:

$$H_c^{\text{aff}} = \left\{ \mathbf{x} = \sum_{k=1}^{n_c} \alpha_{ck} \mathbf{x}_{ck} \,\middle|\, \sum_{k=1}^{n_c} \alpha_{ck} = 1 \right\}, c = 1, \ldots, C. \tag{1}$$

The affine model basically regards any affine combination of an individual's feature sample vectors as a valid face feature sample for that person.

To compute the distance between two affine hulls, we first need to choose a reference point on the affine hull. This reference point can be one of the face image samples of a set or it can be the mean face image of the set. Let the reference point be denoted as $\boldsymbol{\mu}_c$. Then, the affine model of set $c$ in terms of this point is written as:

$$H_c^{\text{aff}} = \left\{ \mathbf{x} = \boldsymbol{\mu}_c + \mathbf{U}_c \mathbf{v}_c \,\middle|\, \mathbf{v}_c \in \mathbb{R}^l \right\}. \tag{2}$$

Here, $\mathbf{U}_c$ is an orthonormal basis for the directions spanned by the affine subspace, $\mathbf{v}_c$ is a vector of free parameters that determines the coordinates for the points within the subspace, expressed with respect to the basis $\mathbf{U}_c$, and $l$ is the number of the basis vectors. Numerically, $\mathbf{U}_c$ is obtained by applying the thin Singular Value Decomposition to $[\mathbf{x}_{c1} - \boldsymbol{\mu}_c, \ldots, \mathbf{x}_{cn_c} - \boldsymbol{\mu}_c]$. Given two non-intersecting affine hulls, $\{\mathbf{U}_c \mathbf{v}_c + \boldsymbol{\mu}_c\}$ and $\{\mathbf{U}_{c'} \mathbf{v}_{c'} + \boldsymbol{\mu}_{c'}\}$, the closest points on them that gives the distance between the affine hulls can be found by solving the following optimization problem:

$$\arg\min_{\mathbf{v}_c, \mathbf{v}_{c'}} ||(\mathbf{U}_c \mathbf{v}_c + \boldsymbol{\mu}_c) - (\mathbf{U}_{c'} \mathbf{v}_{c'} + \boldsymbol{\mu}_{c'})||^2. \tag{3}$$

Defining $\mathbf{U} \equiv (\mathbf{U}_c \; -\mathbf{U}_{c'})$ and $\mathbf{v} \equiv \begin{pmatrix} \mathbf{v}_c \\ \mathbf{v}_{c'} \end{pmatrix}$, this can be written as a standard least squares problem,

$$\arg\min_{\mathbf{v}} \; ||\mathbf{U}\mathbf{v} - (\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_c)||^2, \tag{4}$$

whose solution is $\mathbf{v} = (\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top(\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_c)$. So, the distance between the affine hulls becomes,

$$D(H_c^{\text{aff}}, H_{c'}^{\text{aff}}) = ||(\mathbf{I} - \mathbf{P})(\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'})||, \tag{5}$$

where $\mathbf{P} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top$ is the orthogonal projection matrix of the joint span of the directions contained in the two subspaces.

### 2.2. Generative Convex Hull Models

The convex hull of a set is defined as the smallest convex set containing its samples. When the full affine hull representation given in (1) is restricted for only positive $\alpha_{ck}$ coefficients, it represents the minimal convex set, *i.e.*, the convex hull of the set,

$$H_c^{\text{conv}} = \left\{ \mathbf{x} = \sum_{k=1}^{n_c} \alpha_{ck}\mathbf{x}_{ck} \,\middle|\, \sum_{k=1}^{n_c} \alpha_{ck} = 1, \alpha_{ck} \geq 0 \right\}. \tag{6}$$

Convex hull representation of sets is much tighter than the affine approximation. The distance between two convex hulls can be found by solving the following constrained convex quadratic optimization (QP) problem by using any standard QP solvers:

$$\begin{aligned} (\boldsymbol{\alpha}_c^*, \boldsymbol{\alpha}_{c'}^*) &= \underset{\boldsymbol{\alpha}_c,\, \boldsymbol{\alpha}_{c'}}{\arg\min}\ ||\mathbf{X}_c\,\boldsymbol{\alpha}_c - \mathbf{X}_{c'}\,\boldsymbol{\alpha}_{c'}||^2 \\ \text{s.t.} \quad & \sum_{k=1}^{n_c} \alpha_{ck} = 1 = \sum_{k'=1}^{n_{c'}} \alpha_{c'k'},\ \ \alpha_{ck},\ \alpha_{c'k'} \geq 0. \end{aligned} \tag{7}$$

## 3. Proposed Method

In the proposed method, our goal is to find discriminative and compact affine and convex hull models for each image set such that these models best fit to the samples of their own image sets but are far from image samples of other sets belonging to different people. Generatively learned models explained in the previous section learn an independent model for each person in the gallery by using only the image samples belonging to a particular person of interest. However, we have to consider all data in the gallery to learn a discriminative model for each person, thus we need more efficient algorithms to accomplish this task. In the following, we explain the procedures for finding discriminative affine and convex hull models.

### 3.1. Discriminative Affine Hulls

Assume that we are given an image set belonging to a particular class $c$. Let us denote this class as the positive class and all other remaining classes in the gallery as the negative class. As explained earlier, an affine hull (or an affine subspace) is characterized by basis vectors $\mathbf{U}_+$ and the reference point $\boldsymbol{\mu}_+$, which we choose as the positive class mean, $\boldsymbol{\mu}_+ = \sum_{i=1}^{n_+} \mathbf{x}_i$. Without loss of generality, we can consider an orthonormal basis vector set such that, $\mathbf{U}_+^\top\mathbf{U}_+ = \mathbf{I}$. The distance from any sample $\mathbf{x}$ to this affine hull can be computed by using,

$$d(\mathbf{x}, H_+^{\text{aff}}) = (\mathbf{I} - \mathbf{P}_+)(\mathbf{x} - \boldsymbol{\mu}_+) = \mathbf{P}_+^\perp(\mathbf{x} - \boldsymbol{\mu}_+), \tag{8}$$

where $\mathbf{P}_+ = \mathbf{U}_+\mathbf{U}_+^\top$ is the orthogonal projection operator onto the affine hull of the positive class, and $\mathbf{P}_+^\perp$ is the projection matrix of the orthogonal complement of $\mathbf{P}_+$. It should be noted that orthogonal projection operators are both symmetric and idempotent, i.e., $\mathbf{P}_+^\top = \mathbf{P}_+$ and $\mathbf{P}_+^2 = \mathbf{P}_+$.

Our goal is to find the affine hull that best fits to the positive class samples but at the same time is far as possible from the negative class samples of other image sets. Therefore, we must try to minimize the distances from positive image samples to the affine hull and maximize the distances from negative class samples to the hull. By centering all image samples by using $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \boldsymbol{\mu}_+$, this can be written as the following optimization problem,

$$\begin{aligned} &\underset{\mathbf{U}_+}{\arg\min}\ \tfrac{1}{n_+}\textstyle\sum_{i=1}^{n_+} \left\|\tilde{\mathbf{x}}_i - \mathbf{U}_+\mathbf{U}_+^\top\tilde{\mathbf{x}}_i\right\|^2 - \tfrac{\lambda}{n_-}\sum_{j=1}^{n_-} \left\|\tilde{\mathbf{x}}_j - \mathbf{U}_+\mathbf{U}_+^\top\tilde{\mathbf{x}}_j\right\|^2 \\ &\text{s.t. } \mathbf{U}_+^\top\mathbf{U}_+ = \mathbf{I}. \end{aligned} \tag{9}$$

Here, $n_{+(-)}$ denotes the number of positive (negative) class samples, and $\lambda$ is a parameter that must be set by the user, and it adjusts the weights of distances of negative class samples with respect to the distances of positive class samples. To solve this optimization problem, we first introduce the Lagrangian function given below

$$\begin{aligned} L(\mathbf{U}_+, \boldsymbol{\Lambda}) &= \frac{1}{n_+}\sum_{i=1}^{n_+} \left\|\tilde{\mathbf{x}}_i - \mathbf{U}_+\mathbf{U}_+^\top\tilde{\mathbf{x}}_i\right\|^2 \\ &- \frac{\lambda}{n_-}\sum_{j=1}^{n_-} \left\|\tilde{\mathbf{x}}_j - \mathbf{U}_+\mathbf{U}_+^\top\tilde{\mathbf{x}}_j\right\|^2 + \operatorname{Tr}\boldsymbol{\Lambda}(\mathbf{U}_+^\top\mathbf{U}_+ - \mathbf{I}). \end{aligned} \tag{10}$$

Here, $\boldsymbol{\Lambda} = (\Lambda_{kl})$ denotes the Lagrangian multipliers to enforce the orthonormal constraints on basis vectors. The KKT optimality conditions yield

$$\tfrac{\partial L}{\partial \mathbf{U}_+} = -\tfrac{2}{n_+}\textstyle\sum_{i=1}^{n_+} \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top\mathbf{U}_+ + \tfrac{2\lambda}{n_-}\sum_{j=1}^{n_-} \tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j^\top\mathbf{U}_+ + 2\mathbf{U}_+\boldsymbol{\Lambda} = \mathbf{0}. \tag{11}$$

Let's define a matrix $\mathbf{S}_+$ as

$$\mathbf{S}_+ = \frac{1}{n_+}\sum_{i=1}^{n_+} \tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top - \frac{\lambda}{n_-}\sum_{j=1}^{n_-} \tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j^\top. \tag{12}$$

It should be noted that the matrix $\mathbf{S}_+$ is the difference of two covariance matrices, and hence it is symmetric. It is also positive (semi)-definite for sufficiently small values of $\lambda$.

By inserting this definition in (11), we obtain the following relation,

$$\mathbf{S}_+\mathbf{U}_+ \;=\; \mathbf{U}_+\mathbf{\Lambda}. \tag{13}$$

This is a typical eigen decomposition problem and the solution is the eigenvectors of the symmetric positive (semi)-definite matrix, $\mathbf{S}_+$. It should be noted that the solution is a global minimum, and if $\mathbf{S}_+$ is strictly positive definite then the solution is also a unique global minimum.

This procedure is repeated for each class in the gallery and we obtain a discriminative affine hull model for each class. During online testing, we find the distances from query image samples to each affine hull by using (8) and compute the mean of distances of the $k$ closest samples. Then, the query set is assigned to the class which yields the smallest mean distance. We do not use the distances of all query image samples, instead we use the ones that lie in critical regions where the query set approaches to other gallery sets. The best value of $k$ is determined based on cross-validation.

Lastly, it should be noted that this approach is very different from the Linear Discriminant Analysis methods where a unique embedding space is extracted by using the within-class and between-class scatter matrices. In our setting, there are $C$ different affine subspaces extracted for each person in the gallery, and the decision is made based on the closest distances to the affine subspaces.

### 3.2. Discriminative Convex Hulls

In contrast to the affine hulls, convex hulls have exponentially many facets so they cannot be directly stored explicitly as in affine hulls. Therefore, one has to solve a QP problem online when finding distances from the query samples to the convex hulls of classes in the gallery, which is computationally too expensive. One solution to this problem may be finding the hyperplane that best separates a convex hull of a class from the remaining class samples by using a linear SVM classifier since a linear SVM returns the best separating hyperplane between convex hulls of positive and negative classes [1]. Then, the distances can be approximated by computing the distances from the query samples to the separating hyperplane through simple dot products. However, this solution will not work when the convex hulls of classes in the gallery are not separable by linear hyperplanes as illustrated in Fig. 1. As seen in the figure, it is not possible to separate the convex hull of the red colored class samples in the middle from other classes by a linear hyperplane. On the other hand, polyhedral conic classifier (PCC) of [4] can be used in such cases. As opposed to the linear SVM classifier, PCC classifiers can return tight polyhedral acceptance regions enclosing the positive class samples. These compact polyhedral acceptance regions can be used to approximate the discriminative convex hulls of classes, and computing distances from query samples to the polyhedral acceptance regions is extremely fast since it requires simple dot products as in linear SVMs.

We use the Extended Polyhedral Conic Classifier (EPCC) of [4] to approximate the discriminative convex hull of each class in the gallery. An extended polyhedral conic function of a positive class can be written as,

$$f_{\mathbf{w}_+,\boldsymbol{\gamma}_+,\mathbf{c}_+,b_+}(\mathbf{x}) = \mathbf{w}_+^\top(\mathbf{x}-\mathbf{c}_+)+\boldsymbol{\gamma}_+^\top|\mathbf{x}-\mathbf{c}_+|-b_+, \tag{14}$$

where $\mathbf{x} \in \mathbb{R}^d$ is a test point, $\mathbf{c}_+ \in \mathbb{R}^d$ is the cone vertex, $\mathbf{w}_+ \in \mathbb{R}^d$ is a weight vector, $b_+$ is an offset, $|\mathbf{u}| = (|u_1|, ..., |u_d|)^\top$ denotes the component-wise modulus and $\boldsymbol{\gamma}_+ \in \mathbb{R}^d$ is a corresponding weight vector. We set the cone vertex to the mean of the positive class samples as in [4]. To find a classifier that will return polyhedral acceptance regions, we need to solve the following QP problem for each class in the gallery:

$$\begin{aligned}
\underset{\mathbf{w}_+,\boldsymbol{\gamma}_+}{\arg\min}\;\; & \tfrac{\lambda}{2}\mathbf{w}_+^\top\mathbf{w}_+ + \tfrac{1}{n_+}\textstyle\sum_i \xi_i + \tfrac{1}{n_-}\sum_j \xi_j - \mathbf{s}^\top\boldsymbol{\gamma}_+ \\
\text{s.t.}\;\; & \mathbf{w}_+^\top(\mathbf{x}_i - \mathbf{c}_+) + \boldsymbol{\gamma}_+^\top|\mathbf{x}_i - \mathbf{c}_+| - 1 \le \xi_i,\; i = 1, ..., n_+, \\
& \mathbf{w}_+^\top(\mathbf{x}_j - \mathbf{c}_+) + \boldsymbol{\gamma}_+^\top|\mathbf{x}_j - \mathbf{c}_+| - 1 \ge 1 - \xi_j,\; j = 1, ..., n_-, \\
& \xi_i,\, \xi_j \ge 0.
\end{aligned} \tag{15}$$

Here, $\lambda$ is a regularization weight for $\mathbf{w}_+$, $\mathbf{s} > 0$ is a user-supplied vector of cost penalties for increasing $\boldsymbol{\gamma}_+$, and $b_+$ is fixed to 1. This optimization problem is solved by using Stochastic Gradient (SG) method.

We solve the optimization problem (15) for each class $c$ in the gallery and compute EPCC classifier parameters $\mathbf{w}_c, \boldsymbol{\gamma}_c$. Then, we compute the distances from the query image samples to the polyhedral acceptance regions of each class by using the following function that includes simple dot products,

$$d(\mathbf{x}_{query}, H_c^{\text{conv}}) = \mathbf{w}_c^\top(\mathbf{x}_{query}-\mathbf{c}_c)+\boldsymbol{\gamma}_c^\top|\mathbf{x}_{query}-\mathbf{c}_c|-1. \tag{16}$$

As in the affine hull case, we compute the $k$ closest query sample distances to each gallery class and compute their mean, $\bar{\mathbf{x}}_c$. Then, we assign the query set, $\mathbf{X}_q$, to the class that yields the smallest distance, i.e., we use the following decision function,

$$g(\mathbf{X}_q) \;=\; \min_{c=1,...,C}(\bar{\mathbf{x}}_c). \tag{17}$$

The online decision process is extremely fast even compared to the linear affine hulls since we need to implement two simple dot products. For discriminative affine hulls, the number of eigenvectors spanning the discriminative hulls is mostly larger than 2. Therefore, the decision process is much slower compared to the discriminative convex hulls, which are approximated by using compact polyhedral acceptance regions.

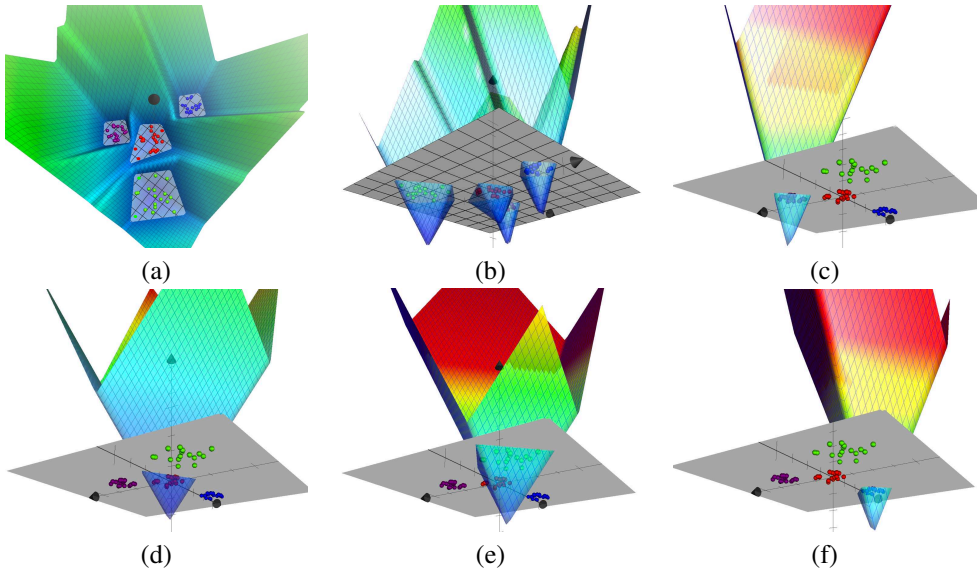

Figure 1. Visualization of PCC classifiers for 2D synthetic data: The positive acceptance regions of classes are "kite-like" octahedroids containing the corresponding class points for which a linear hyperplane lies above an $L_1$ cone. (a): 2D data samples for 4 different classes and the resulting approximated convex hulls. (b): the combined polyhedral conic functions of classes in 3D. (c)-(f): views of positive-class acceptance regions in 3D for each class.

It should be noted that Cevikalp and Yavuz [5] also used EPCC classifiers for set based recognition. However, their setup is completely different than the one we propose here. Both Cevikalp and Yavuz [5] and Hayat et al. [14] use discriminative classifiers for separation of image sets. But, they train to separate the query set images from the gallery images, and this requires an online training of a binary classifier during testing. As a result, the test times of these two methods are very long for real-time applications especially for large-scale datasets as given in our experiments. In contrast, our learning process is accomplished offline and we just implement simple dot products during online testing.

## 4. Experiments

We tested the proposed methods, Discriminative AH and Discriminative CH, using discriminative affine/convex hulls on 5 datasets used for both image set based face recognition and verification. These are Point-and-Shoot Face Recognition Challenge (PaSC) [2], YouTube Celebrities [21], COX [17], ESOGU [33], and FaceScrub[25] datasets. The images are represented by using either Local Binary Pattern (LBP) features or Convolutional Neural Network (CNN) features. We compared the proposed methods to other methods using generatively learned affine/convex hulls as well as some different models in the literature including convex hull method (CHISD) [3], affine hull method (AHISD) [3], Binary EPCC [5], SANP [16], Mutual Subspace Method (MSM) [34], Regularized Nearest Points (RNP) [36], SPD Manifolds [18], Manifold-Manifold Distance (MMD) [29], Collaboratively Regularized Nearest Points(CRNP) [32] and Self-Regularized Non-negative Adaptive Distance Metric Learning (SRN-ADML) [24]. For SPD Manifolds, the covariance matrices of sets augmented with their mean are used to create SPD matrices as in [18], and we use the Log-Euclidean metric to measure the similarities between them.

### 4.1. Face Verification Experiments on PaSC Dataset

For face verification experiments, we used Point-and-Shoot Face Recognition Challenge (PaSC) dataset [2]. The PaSC dataset includes 2802 videos of 265 people carrying out simple actions. Videos are recorded under two different settings. In our experiments, we used deep CNN features of face images provided by [19]. On PaSC, there are two video face verification experiments: control-to-control and handheld-to-handheld experiments. In both experiments, the target and query sets contain the same set of videos. The task is to verify a claimed identity in the query video by comparing with the associated target video. Since the same 1401 videos served as both the target and query sets, "same video" comparisons are excluded as in [19], and our results are directly comparable to the ones reported in [19] since we use the same CNN features and test protocols.

To test methods, we follow the same testing setup as used in [19]: we first compute the similarities between pair-wise face videos and create a similarity matrix. Then, this matrix is used to create ROC curves and we report the verification rate when false accept rate is 0.01. In addition, we also report the average precision (mAP) scores obtained from Precision-Recall curves. We set the number of closest query

Table 1. Results for PaSC Face Verification Experiments.

| Methods | PaSC Control | | PaSC Handheld | |
|---|---|---|---|---|
| | **Verification Rates(%)** | **mAP (%)** | **Verification Rates(%)** | **mAP (%)** |
| Discriminative AH | 88.76 | 70.66 | 77.39 | 60.37 |
| Discriminative CH | **93.31** | **76.82** | **85.00** | **68.33** |
| Linear AHISD | 80.97 | 61.90 | 65.14 | 46.76 |
| Linear CHISD | 88.48 | 70.11 | 74.01 | 55.49 |
| MSM | 84.16 | 65.23 | 69.68 | 50.55 |
| SANP | 86.88 | 67.48 | 71.73 | 53.12 |
| RNP | 75.51 | 54.62 | 54.84 | 36.42 |
| SRN-ADML | 84.63 | 66.93 | 69.36 | 52.10 |
| CERML-EG [19] | 80.11 | – | 77.37 | – |
| DAN [27] | 92.06 | – | 80.33 | – |

samples $k$ to 20 for the Discriminative AH method and to 10 for the Discriminative CH method. The results are reported in Table 1, and they support our claim that discriminative models significantly outperform the generatively learned models. More specifically, the proposed Discriminative AH method brings around 8% improvement over generatively learned affine hulls on control dataset and it brings around 12% improvement on handheld dataset. Similarly, the proposed Discriminative CH method improves its generative counterpart around 5% on control dataset and around 11% on handheld dataset. Moreover, the proposed Discriminative CH method achieves the best accuracy among all tested methods and it significantly outperforms the previous state-of-the-art results of DAN (discriminative aggregations network) method [27] as well as accuracies of CERML-EG [19] using the same CNN features we used in our tests. To the best of our knowledge, our results are the best accuracies reported on PaSC dataset in the literature.

## 4.2. Experiments on Set Based Face Recognition

### 4.2.1 Experiments on the YouTube Celebrities Dataset

The YouTube Celebrities dataset contains 1910 videos of 47 celebrities that are collected from YouTube. Each sequence includes different number of frames that are mostly low resolution. The dataset does not provide the cropped faces from videos. Therefore, we manually cropped faces using a semi-automatic annotation tool and resized them to 40×40 gray-scale images. Then, we extracted LBP features. We conduct 10 runs of experiments by randomly selecting 9 videos (3 for training, 6 for testing) for each experiment by following the same protocol of [37, 28].

The averages of the classification rates and testing times are shown in Table 2. The number of the nearest query samples $k$ is set to 7 for both discriminative affine and convex hull methods. Both proposed methods significantly outperform their classical counterparts using generative affine/convex hulls, but their accuracies are slightly behind the Binary EPCC which achieves the highest accuracy. However, the proposed methods are the most efficient methods in terms of testing time. For example, the proposed Discriminative CH method is approximately 57 times faster than the Binary EPCC whereas its accuracy is only 0.9% behind the Binary EPCC.

Table 2. Classification Rates (%) and Testing Times on the YouTube Celebrities Dataset.

| Method | Accuracy | Testing Time |
|---|---|---|
| Discriminative AH | $71.5 \pm 2.0$ | 0.3 sec |
| Discriminative CH | $71.2 \pm 2.2$ | **0.1** sec |
| Linear AHISD | $62.0 \pm 2.4$ | 39.7 sec |
| Linear CHISD | $65.7 \pm 2.5$ | 23.6 sec |
| Binary EPCC | $\mathbf{72.1} \pm 2.4$ | 5.7 sec |
| MSM | $63.8 \pm 2.3$ | 2.4 sec |
| SANP | $58.6 \pm 3.8$ | 75.5 sec |
| RNP | $68.4 \pm 2.6$ | 20.6 sec |
| CRNP | $71.2 \pm 2.6$ | 76.6 sec |
| SPD Manifolds | $58.9 \pm 2.3$ | 7.1 sec |
| SRN-ADML | $63.6 \pm 2.9$ | 76.1 sec |
| MMD | $65.5 \pm 2.5$ | 34.0 sec |

### 4.2.2 Experiments on the ESOGU-285 Face Videos Dataset

The ESOGU-285 database [33, 6] is a video dataset which consists of 285 people with 8 videos for each person. Videos are captured in an indoor environment in two separate sessions under four different scenarios. The total num-

Table 3. Classification Rates (%) and Testing Times on the ESOGU-285 Video Dataset.

| Methods | LBP Features | | CNN Features | |
|---|---|---|---|---|
| | **Accuracy** | **Testing Time** | **Accuracy** | **Testing Time** |
| Discriminative AH | 86.6 | 3.7 sec | 85.26 | 10.6 sec |
| Discriminative CH | **89.0** | **2.8** sec | **85.70** | **7.6** sec |
| Linear AHISD | 66.8 | 180.0 sec | 81.32 | 543.8 sec |
| Linear CHISD | 76.6 | 390 sec | 79.82 | 378.6 sec |
| Binary EPCC | 86.4 | 250.5 sec | 81.32 | 2268.9 sec |
| MSM | 69.6 | 5.1 sec | 77.02 | 5.6 sec |
| SANP | 79.1 | 564.6 sec | 81.66 | 1087.6 sec |
| RNP | 51.9 | 2205.3 sec | 80.79 | 367.8 sec |
| CRNP | OOM | $--$ | OOM | $--$ |
| SPD Manifolds | 64.65 | 56.5 sec | 76.31 | 63.6 sec |
| SRN-ADML | 68.4 | 380.2 sec | 77.46 | 458.5 sec |
| MMD | 77.6 | 150.4 sec | 79.04 | 28.9 sec |

ber of the frames is 764006 in 2280 video sequences. This is the largest dataset used in this study in terms of the total number of frames. We used both LBP and CNN features of image samples. To extract CNN features, we used the recent state-of-the-art ResNet-101 architecture [15].

In our experiments, we used the first session videos to form the gallery sets and the second session videos were used in testing as in [5]. Experimental results are given in Table 3. We could not implement CRNP because of memory issues and "OOM" in the table indicates the "out of memory" problem. The number of the nearest query samples, $k$, is set to 95 for both affine and convex hulls. As seen in the table, the proposed methods achieve the best accuracies for both LBP and CNN features, and they significantly outperform all other tested methods. The proposed Discriminative AH method outperforms Linear AHISD by 19.8% for LBP features and by 3.9% for CNNs. Similarly, Discriminative CH method beats Linear CHISD method by 12.4% for LBPs and by 5.9% for CNNs. These results clearly demonstrate the superiority of discriminative models over the generative ones. Moreover, to the best of our knowledge, the accuracy of the proposed Discriminative CH method, 89.0%, is the best accuracy reported in literature on this dataset. The proposed methods are also the most efficient methods in terms of testing time. Another discriminant classifier, the Binary EPCC, achieves the third best accuracy for LBP features; but it is very slow compared to the proposed methods. More precisely, our Discriminative CH method is approximately 89 times faster than the Binary EPCC for LBP features and it is 298 times faster with CNN features. Similarly, the proposed Discriminative AH method is approximately 68 times faster than the Binary EPCC for LBPs and it is 214 times faster for CNN features. It is also worth mentioning that, the performances of all generative methods significantly improve over LBP features when we use ResNet-101 CNN features. This is very natural since the discriminative information is already included in learned CNN features no matter how we train generative methods. But, there is no improvement in discriminative methods. In fact, their accuracies are lower compared to LBP features which is very unexpected. This clearly shows that the accuracy of classical deep neural network based methods trained with single images can be improved for set based recognition where the images have different poses including full left/right profile views in addition to the frontal views. To this end, we must train such nets with image sets and enforce to minimize the distances between the different pose image features in the same set (e.g., by using triplet loss function instead of common softmax loss) to obtain higher accuracies.

#### 4.2.3 Experiments on the COX Video to Video Dataset

The COX Faces dataset contains 3000 video sequences of 1000 walking individuals [17]. The videos are captured with three fixed camcorders while the subjects walk around the pre-designed S-shape route. For this database, we used LBP features extracted from $32\times40$ histogram equalized face images. We did not extract CNN features because of the small size of the images. There are 3 image sets per person. We chose one set from each person for testing, and the remaining two sets were used as gallery. For the second and the third trials, we have chosen the test set from the sets that were not used for testing earlier.

The classification rates are the averages of these three trials, and they are given in Table 4. The number of the nearest query samples $k$ is set to 20 for both discriminative affine and convex hull methods. Similar to the ESOGU dataset, the proposed methods again significantly outperform the methods using generative affine and convex hulls.

The proposed Discriminative AH method improves the accuracy by 11.6% over Linear AHISD, and the Discriminative CH method improves the accuracy by 30.3% over the Linear CHISD. The best accuracy is obtained by the proposed Discriminative CH method among all tested methods, and it significantly outperforms discriminative Binary EPCC method of [5]. The accuracy improvement is approximately 11%, which is quite large. The proposed methods are also the most efficient methods in terms of testing times. For example, the proposed Discriminative CH method is approximately 100 times faster than the discriminative Binary EPCC method.

Table 4. Classification Rates (%) and Testing Times on the COX Video Dataset.

| Method | Accuracy | Testing Time |
|---|---|---|
| Discriminative AH | $55.9 \pm 13.6$ | 3.4 sec |
| Discriminative CH | $\mathbf{75.1} \pm 1.6$ | **1.7** sec |
| Linear AHISD | $44.3 \pm 9.8$ | 82.9 sec |
| Linear CHISD | $44.8 \pm 11.3$ | 54.3 sec |
| Binary EPCC | $64.0 \pm 11.5$ | 171.7 sec |
| MSM | $41.6 \pm 5.3$ | 18.6 sec |
| SANP | $43.6 \pm 11.2$ | 978.7 sec |
| RNP | $45.4 \pm 13.7$ | 217.3 sec |
| CRNP | OOM | $--$ |
| SPD Manifolds | $33.1 \pm 9.1$ | 97.3 sec |
| SRN-ADML | $44.6 \pm 7.9$ | 351.7 sec |
| MMD | $42.7 \pm 10.5$ | 60.3 sec |

#### 4.2.4 Experiments on the FaceScrub Dataset

The FaceScrub dataset [25] includes face images of 530 celebrities. It has been created by detecting faces based on automated search of public figures on the internet followed by manually checking and cleaning the results. In the dataset, there are 265 male and 265 female celebrities' face images. We manually checked the face images and cleaned non-face images since there were still some annotation mistakes. As a result, we had 67,437 face images of 530 celebrities with an average of 127 images (minimum 39, maximum 201) per person which is suitable to form image sets. The face images are mostly high resolution frontal face images and we resized them to $128 \times 128$. We extracted CNN features of these images.

In our tests, we first divided the dataset into 4 equal folds, and we used the images of one fold as the gallery and the remaining images are used for testing (i.e. 530 image sets are used as the gallery and the remaining $3 \times 530 = 1590$ image sets are used as the test set). This is repeated 4 times for each fold and the final accuracy is the average of the results obtained in each trial. The number of the nearest query samples, $k$, is set to 7 for both affine and convex hulls. The accuracies and test times of the compared methods are given in Table 5. As can be seen in the table, all tested methods achieve very high accuracies around 100%. The proposed methods again achieve the highest accuracies. The proposed methods are also faster than all tested methods with the exception of MSM method, which is the most efficient method in terms of testing time for this dataset.

Table 5. Classification Rates (%) and Testing Times on the FaceScrub Dataset.

| Method | Accuracy | Testing Time |
|---|---|---|
| Discriminative AH | $\mathbf{100} \pm 0.00$ | 1.20 sec |
| Discriminative CH | $\mathbf{100} \pm 0.00$ | 0.70 sec |
| Linear AHISD | $99.94 \pm 0.05$ | 6.17 sec |
| Linear CHISD | $99.97 \pm 0.04$ | 8.33 sec |
| Binary EPCC | $\mathbf{100} \pm 0.00$ | 46.2 sec |
| MSM | $99.94 \pm 0.05$ | **0.30** sec |
| SANP | $99.94 \pm 0.05$ | 75.40 sec |
| RNP | $\mathbf{100} \pm 0.00$ | 10.81 sec |
| CRNP | OOM | $--$ |
| SPD Manifolds | $96.20 \pm 3.3$ | 19.5 sec |
| SRN-ADML | $99.95 \pm 0.04$ | 14.78 sec |
| MMD | $\mathbf{100} \pm 0.00$ | 2.45 sec |

## 5. Conclusion

This paper introduces discriminative affine/convex hulls for image set based face recognition. As opposed to the other methods that learn the generative models approximating face image sets independently, the proposed methods learn discriminative models by incorporating all image sets belonging to different people in the gallery. As a result, the accuracies are significantly improved over the methods that use generatively learned models. The proposed methods also significantly outperform the discriminative methods used for set based recognition. The accuracy improvement is very significant especially on the challenging PaSC, ESOGU and COX datasets. For example, the proposed Discriminative CH method outperforms another successful discriminative method, Binary EPCC of [5] on COX dataset by 11.1%, which is quite significant. All these results verify that the proposed discriminative models are better suited than generative models for set based classification. In addition to these accuracy gains, the proposed methods are also extremely fast since we learn discriminative model parameters offline and we just implement simple matrix multiplications during online testing. As a result, we obtained speed-ups to a factor of 298 over other discriminative methods in the literature.

**Acknowledgments:** This work was supported by the Scientific and Technological Research Council of Turkey (TUBİTAK) under grant number EEEAG-118E294.

## References


[1] Kristin P. Bennett and Erin J. Bredensteiner. Duality and geometry in svm classifiers. In *International Conference on Machine Learning*, 2000. 4

[2] Ross Beveridge, Jonathon Phillips, David S. Bolme, Bruce A. Draper, Geof H. Givens, Yui Man Lui, Mohammad Nayeem Teli, Hao Zhang, Todd Scruggs, Kevin W. Bowyer, Patrick J. Flynn, and Su Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *BTAS*, 2013. 5

[3] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010. 1, 2, 5

[4] Hakan Cevikalp and Bill Triggs. Polyhedral conic classifiers for visual object detection and classification. In *CVPR*, 2017. 4

[5] Hakan Cevikalp and Hasan Serhan Yavuz. Fast and accurate face recognition with image sets. In *ICCV Workshops*, 2017. 2, 5, 7, 8

[6] Hakan Cevikalp, Hasan Serhan Yavuz, and Bill Triggs. Face recognition based on videos by using convex hulls. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2019. 6

[7] Shaokang Chen, Conrad Sanderson, Mehrtash T. Harandi, and Brian C. Lovell. Improved image set classification via joint sparse approximated nearest subspaces. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013. 1

[8] Yi-Chen Chen, Vishal M. Patel, Sumit Shekhar, Rama Chellappa, and P. Jonathon Phillips. Video-based face recognition via joint sparse representation. In *IEEE Int. Conf. on Automat. Face Gesture Recognit.*, 2013. 1

[9] Zhen Cui, Hong Chang, Shiguan Shan, Bingpeng Ma, and Xilin Chen. Joint sparse representation for video-based face recognition. *Neurocomputing*, 135:306–312, 2014. 1

[10] Jihun Hamm and Daniel D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Int. Conf. Mach. Learn.*, 2008. 1

[11] Munawar Hayat, Mohammed Bennamoun, and Senjian An. Learning non-linear reconstruction models for image set classification. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014. 2

[12] Munawar Hayat, Mohammed Bennamoun, and Senjian An. Reverse training: an efficient approach for image set classification. In *ECCV*, 2014. 2

[13] Munawar Hayat, Mohammed Bennamoun, and Senjian An. Deep reconstruction models for image set classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37:713–727, 2015. 2

[14] Munawar Hayat, Salman H. Khan, and Mohammed Bennamoun. Empowering simple binary classifiers for image set based face recognition. *International Journal of Computer Vision*, 123:479–498, 2017. 2, 5

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[16] Yiqun Hu, Ajmal S. Mian, and Robyn Owens. Face recognition using sparse approximated nearest points between image sets. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):1992–2004, 2012. 1, 5

[17] Zhiwu Huang, Shiguang Shan, Ruiping Wang, Haihong Zhang, Shihong Lao, Alifu Kuerban, and Xilin Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Transactions on Image Processing*, 24:5967–5981, 2015. 5, 7

[18] Zhiwu Huang, Ruiping Wang, Xianqiu Li, Wenxian Liu, Shiguang Shan, Luc Van Gool, and Xilin Chen. Geometry-aware similarity learning on SPD manifolds for visual recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28:1318–1322, 2018. 1, 5

[19] Zhiwu Huang, Ruiping Wang, Shiguang Shan, Luc Van Gool, and Xilin Chen. Cross euclidean-to-riemannian metric learning with application to face recognition from video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2827–2840, 2018. 1, 5, 6

[20] Zhiwu Huang, Ruiping Wang, Shiguang Shan, Xianqiu Li, and Xilin Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *ICML*, 2015. 1

[21] Minyoung Kim, Sanjiv Kumar, Vladimir Pavlovic, and Henry Rowley. Face tracking and recognition with visual constraints in real-world videos. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008. 5

[22] Luoqi Liu, Li Zhang, Hairong Liu, and Shuicheng Yan. Toward large-population face identification in unconstrained videos. *IEEE Trans. on Circuits and Systems for Video Technology*, 24:1874–1884, 2014. 2

[23] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017. 2

[24] Ajmal Mian, Yiqun Hu, Richard Hartley, and Robyn Owens. Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning. *IEEE Transactions on Image Processing*, 22:5252–5262, 2013. 5

[25] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *IEEE International Conference on Image Processing*, 2014. 5, 8

[26] Yongming Rao, Ji Lin, Jiwen Lu, and Jie Zhou. Learning discriminative aggregation network for video-based face recognition. In *IEEE Conference on Computer Vision*, 2017. 2

[27] Yongming Rao, Ji Lin, Jiwen Lu, and Jie Zhou. Learning discriminative aggregations network for video-based face recognition. In *ICCV*, 2017. 6

[28] Ruiping Wang and Xilin Chen. Manifold discriminant analysis. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009. 6

[29] Ruiping Wang, Shiguang Shan, Xilin Chen, and Wen Gao. Manifold-manifold distance with application to face recognition based on image sets. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008. 5

[30] Tiesheng Wang and Pengfei Shi. Kernel grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognition Letters*, 30:1161–1165, 2009. 1

[31] Wen Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Prototype discriminative learning for image set classification. *IEEE Signal Processing Letters*, 24:1318–1322, 2017. 1

[32] Yang Wu, Michihiko Minoh, and Masayuki Mukunoki. Collaboratively regularized nearest points for set based recognition. In *British Machine Vision Conference*, 2013. 1, 5

[33] Meltem Yalcin, Hakan Cevikalp, and Hasan Serhan Yavuz. Towards large-scale face recognition based on videos. In *International Conference on Computer Vision Workshop on Video Summarization for Large-Scale Analytics*, 2015. 5, 6

[34] Osamu Yamaguchi, Kazuhiro Fukui, and Ken ichi Maeda. Face recognition using temporal image sequence. In *International Symposium of Robotics Research*, pages 318–323, 1998. 1, 5

[35] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Fang Wen Dong Chen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017. 2

[36] Meng Yang, Pengfei Zhu, Luc Van Gool, and Lei Zhang. Face recognition based on regularized nearest points between image sets. In *IEEE Int. Conf. on Automat. Face Gesture Recognit.*, 2013. 1, 5

[37] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Simon Chi-Keung Shiu, and David Zhang. Image set-based collaborative representation for face recognition. *IEEE Transactions on Information Forensics and Security*, 9:1120–1132, 2014. 1, 6