IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 30, 2017, accepted January 9, 2018, date of publication January 26, 2018, date of current version March 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2794357

Multi-Order Statistical Descriptors for Real-Time Face Recognition and Object Classification

ARIF MAHMOOD[1], MUHAMMAD UZAIR[2], AND SOMAYA AL-MAADEED[1]
[1]Department of Computer Science and Engineering, Qatar University, Doha, Qatar
[2]COMSAT Institute of Technology at Wah, Wah Cant, Pakistan

Corresponding author: Arif Mahmood (arif.mahmood@qu.edu.qa)

This work was supported by NPRP through the Qatar National Research Fund (a member of Qatar Foundation) under Grant 7-1711-1-312.

ABSTRACT We propose novel multi-order statistical descriptors which can be used for high speed object classification or face recognition from videos or image sets. We represent each gallery set with a global second-order statistic which captures correlated global variations in all feature directions as well as the common set structure. A lightweight descriptor is then constructed by efficiently compacting the second-order statistic using Cholesky decomposition. We then enrich the descriptor with the first-order statistic of the gallery set to further enhance the representation power. By projecting the descriptor into a low-dimensional discriminant subspace, we obtain further dimensionality reduction, while the discrimination power of the proposed representation is still preserved. Therefore, our method represents a complex image set by a single descriptor having significantly reduced dimensionality. We apply the proposed algorithm on image set and video-based face and periocular biometric identification, object category recognition, and hand gesture recognition. Experiments on six benchmark data sets validate that the proposed method achieves significantly better classification accuracy with lower computational complexity than the existing techniques. The proposed compact representations can be used for real-time object classification and face recognition in videos.

INDEX TERMS Face recognition, image set classification, covariance features, dimensionality reduction.

I. INTRODUCTION

Classification using image sets has recently received significant research attention from the computer vision community [1]–[14]. In image set classification, classifiers are trained using representations learned from one or more image sets containing arbitrary number of images to model a class. This is different from the traditional single image based classification scheme where single image representations are used to train classifiers. The test stage involves assigning a label to a query image set by maximizing a suitable similarity index between gallery and test image set representations. Such type of learning takes advantages from the availability of complementary within class image variations, such as non-rigid deformations, scale, pose and illumination variations, offered by multiple images in a set. Therefore, compared to single mug-shot based face recognition or object categorization in general [15]–[17], set based modeling offers significantly more accurate recognition results.

Image set based classification is naturally applicable to many challenging computer vision problems such as video sequence face recognition [2], [18], action and activity recognition [19]–[22], video surveillance [23], person re-identification in a network of cameras [24] and long-term observations based classification [25], [26]. In general image-set classification is also applicable to the scenarios where the images in a set have significant variations without necessarily having a strict temporal relationship [10].

Classification based on image sets is usually a two-step process. The first step involves learning robust image set representations by efficiently encoding the within set intra and inter-sample dynamics. The second step is concerned with defining a similarity or distance metric for comparing two set representations. A classifier is then trained using the set representations and comparison metric(s). Usually, the classification accuracy, computational efficiency and memory requirements are strongly dependent on both the set representation approach as well as the set to set similarity/distance metric.

Learning efficient, compact and discriminative image set representations is a challenging task. Current image set

representation methods which are relatively more accurate have higher computational burden [10]. On the other hand, the more faster algorithms compromise on the accuracy and robustness. In contrast, we propose an image set representation technique that is compact (memory efficient), accurate and computationally very efficient. This is verified by our experimental results on six benchmark datasets involving different image set based classification tasks.

A. RELATED WORK

We roughly categorize the existing literature on image-set classification into sample based set representation and structure based set representations methods.

The sample based methods rely on the similarity/differences of individual set samples to define a set to set similarity/distance index in a nearest neighbour based approach. These methods also usually generate new samples from the existing ones using approximations techniques. For example, Cevikalp and Triggs [9] proposed to model image-sets as dense convex geometric regions in the feature space. They represented image sets using their affine and convex hull approximations. Gallery and probe sets were then matched using the minimum distance between their affine (AHISD) or convex (CHISD) hulls. For this purpose, linear least squares and SVM based formulations were adopted. Hu *et al.* [10] proposed to generate new intermediate sample representations by using a sparse set of samples from the two sets that are being matched. These intermediate sparse approximated nearest points (SANP) were computed such that they lie on the facet which is closest to affine/convex hulls of the similar sets. Set classification was then performed by simple nearest neighbour based approaches using the SANPs. Similarly, Yang *et al.* [27] used ℓ_2 regularized affine hulls to represent sets and computed regularized nearest points (RNP) from these representations. RNP is shown to be computationally efficient than SANP. Due to their inherent sample matching based mechanism, the accuracy of the sample based techniques is strongly affected by the presence of outlier samples in a set. Furthermore, the sample based algorithms also have relatively higher computational cost due to the approximation process and the constraints imposed on the nearest sample-based matching mechanism.

The structure based methods represent image sets using set structure computed from the linear subspaces [28], [29], mixture of subspaces [13], [30] and non-linear manifolds [7], [8], [11], [31], [32]. Set matching is done using the structural similarity measures such as subspace to subspace distance [29], manifold to manifold distance [8], [31] or by defining kernel function on the manifolds [11], [32], [33] for distance computation. For example, Kim *et al.* [29] represented an image set using linear subspace learned with PCA and formulated set matching as a discriminative learning problem using canonical correlations between the sets. Wang *et al.* [31] used multiple linear subspaces to represent a single image set. Set samples were divided into disjoint and correlated clusters and each cluster is then represented using linear subspace.

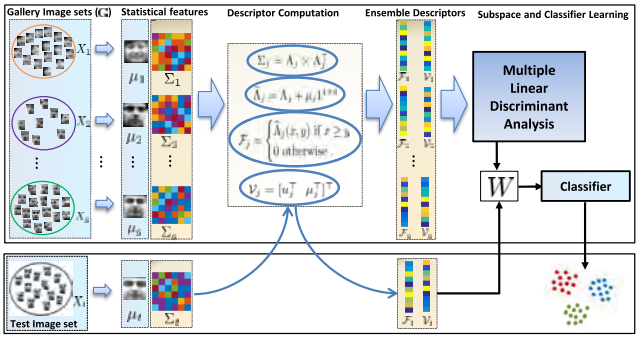

FIGURE 1. An overview of the proposed algorithm including the feature extraction, learning the discriminative basis, training the classifiers and testing the probe sets.

Canonical correlation similarity measure is then used as a similarity index in a nearest neighbour based framework for set matching. Wang and Chen [8] also used multiple linear subspaces to represent a single image set. However, they performed discriminative learning on these representations for improved accuracy. Set matching is then performed in the embedded discriminative space. Harandi *et al.* [32] used Grassmannian manifolds theory and presented image sets as points on the Grassman manifold (linear subspaces). Set matching is then formulated as a discriminative kernel based classifier learning problem by defining suitable Grassmann kernels. Similarly, Wang *et al.* [11] used theory from Reimannian manifolds for image set representation and classification. They used covariance features for image set representation and performed discriminative learning on the Reimannian kernels computed from the set representations for set matching. Structure based methods are efficient, however, they need large number of samples in image sets for better structure estimation. The accuracy of structure based techniques usually depends on the structure estimation methods used. Another factor is the number of images available in a set for detailed and correct structure learning.

In the current manuscript, we extend image set classification algorithms by proposing a new image-set representation which is more accurate under certain conditions. In addition, we extend classification technique to incorporate an efficient and more accurate Kernel Linear Discriminant Analysis (KLDA) as classifier. KLDA has consistently exhibited more accuracy over MLDA. The proposed algorithm is shown in Fig. 1. We also perform comprehensive experiments to test the robustness of the proposed algorithms to noisy image-sets and to the presence of outliers in the image-sets. Furthermore, our implementation of the proposed algorithm has been optimized to yield significantly more speedup than previously reported results. The proposed algorithm is 21, 42 and 716 times faster than CDL [11], DCC [29] and SANP [10] approaches while consistently achieving more accuracy than all of these algorithms.

Due to the inherent simplicity, high accuracy and significant speedup, the proposed algorithm may become a baseline for the performance comparison of the future

research in image-set classification. Some preliminary results were published in [12]. In this paper, two more types of image-set representations and classifiers are proposed which makes four variants. We compare the recognition rate (classification accuracy) and the execution time of the proposed algorithms with ten state of the art image-set classification algorithms. Experimental results demonstrate that the proposed algorithm can achieve significantly better classification accuracy at lower computational complexity than the existing techniques.

II. PROPOSED ALGORITHM

Given training image sets, the proposed algorithm uses a three-step process to build an image set classifier. In the first step, we represent image sets using an efficient and compact feature vector. In the second step, we learn a discriminative subspace and project the feature vectors into that space to achieve dimensionality reduction. Finally, a classifier is trained in the learned discriminative low dimensional space. This increases accuracy and at the same time reduces memory consumption and CPU time. In the test stage, a probe image set is first represented using our proposed feature representation and then projected to the discriminative subspace where it is classified into one of the gallery classes by the learned classifier.

A. MODELING IMAGE SETS AND DISTANCE MEASUREMENTS

An image $x \in \mathcal{R}^d$ with d dimensional feature representation may be considered as a point in a d dimensional space and an image-set is then a point-cloud in that space. We hypothesize that the point-cloud of a given class has its unique properties that can be modeled to uniquely represent that class. If the point-cloud follows a multivariate Gaussian distribution with the following density function

$$P(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}[x-\mu]^\top \Sigma^{-1}[x-\mu]}. \quad (1)$$

Instead of evaluating $P(x|\mu, \Sigma)$, Mahalanobis distance may also be used to compute the distance of an image-set y from a Gaussian distribution as

$$\Delta_M(y, \mu, \Sigma) = [y-\mu]^\top \Sigma^{-1}[y-\mu]. \quad (2)$$

Similarly, the distance between two point clouds having Gaussian distributions can be computed using the Kullback-Leibler (KL) divergence

$$\Delta_{Kl}(P_j||P_i) = \frac{1}{2}\log\frac{|\Sigma_i|}{|\Sigma_j|} + \frac{1}{2}Tr(\Upsilon_{i|j}) - \frac{d}{2}, \quad (3)$$

where $\Upsilon_{i|j} = \Sigma_j\Sigma_i^{-1} + \Sigma_i^{-1}(\mu_j - \mu_i)(\mu_j - \mu_i)^\top$. Since KL divergence is asymmetric, it is not a metric: $\Delta_{Kl}(P_j||P_i) \neq \Delta_{Kl}(P_i||P_j)$. Moreover, KL divergence does not follow the triangular inequality.

Another method to compute distance between two point clouds is to compare their sample covariance matrices. However, these types of representations do not follow the Euclidean geometry but exist rather on a Riemannian manifold. Therefore, methods from Reimannian geometry are used for computing distances such as the affine invariant distance $\Delta_\parallel$ [34]

$$\Delta_\parallel(\Sigma_i, \Sigma_j) = \sqrt{\sum_{p=1}^{d} \ln^2 \lambda_p(\Sigma_i, \Sigma_j)}, \quad (4)$$

where $\lambda_p(\Sigma_i, \Sigma_j)$ denotes the Eigenvalues of the system of equations obtained from the determinant $|\lambda\Sigma_i - \Sigma_j| = 0$. $\Delta_\parallel$ is a metric on the Riemannian manifold over the space $Sym^+(d, \mathbb{R})$ of real semi-positive definite matrices. Moreover, it is also invariant to affine transformations and inversions.

Another metric for such comparisons is the the log Euclidean distance measure Δ_ℓ [35]

$$\Delta_\ell(\Sigma_i, \Sigma_j) = ||\log(\Sigma_i\Sigma_j^{-1})||_F, \quad (5)$$

where $||\cdot||_F$ denotes the Frobenius norm of a matrix. Given the eigen decomposition of a SPD matrix Σ as $\Sigma = U\Lambda U^\top$, we can define its logarithm by $\log\Sigma = U\log(\Lambda)U^\top$. Note that both $\Delta_\parallel$ and Δ_ℓ have high computational cost due to the computation of exponential and logarithm for the high dimensional covariance matrices. Moreover, these methods ignore the important parameters μ_i and μ_j in distance computation.

For images spanning $\mathcal{R}^d$, the corresponding Riemannian manifold will span $\mathcal{R}^{d^2}$, which grows exponential with the growth of image size. For example, images in $\mathcal{R}^{100\times100}$, the corresponding covariance matrix will span $\mathcal{R}^{10^8}$ space. The covariance matrices are symmetric, therefore has $\frac{d(d+1)}{2}$ unique elements. Instead of directly extracting these unique elements, we propose to apply Cholesky decomposition [36] and compute a lower triangular matrix capturing the structure and information of a regularized covariance matrix. The choice of Cholesky decomposition is motivated by the work of Hong *et al.* [37]. They showed that Cholesky decomposition is more efficient for distance computation compared to the previously used affine-invariant or log-Euclidean Riemannian metrics.

Subsequently, we enrich the compressed second order statistic Σ with the first order statistic μ, using two feature level fusion approaches and thus we compute two different descriptors for the image-sets. In the first approach, we add the two statistics to obtain a compressed multi-order statistical descriptor representing an image-set. The dimensionality of this descriptor is $d(d+1)/2$. In the second approach, we concatenate the first order statistic with the lower triangular matrix to obtain the second type of multi-order statistical descriptor. The dimensionality of this descriptor is $d(d+3)/2$. The dimensionality of the both proposed descriptors is further reduced by using Multiple Linear Discriminant Analysis (MLDA) and Kernel LDA (KLDA). At the test time, for each test image set we compute both proposed descriptors and transform these descriptors by the MLDA and KLDA basis learned over the training sets. The LDA space is linear

therefore linear classifiers, such as SVM, are trained to discriminate one class from the others. We observe that in LDA space nearest neighbour classifier also performs excellent. The performance of linear classifiers is leveraged by the fact that the learned representations are linear and discriminative.

B. THE PROPOSED MULTI-ORDER STATISTICAL DESCRIPTORS

Let $G = \{X_j\}_{j=1}^{g} \in \mathcal{R}^{d\times N}$ be the gallery containing g image sets and N is the total number of images such that $N = \sum_{j=1}^{g} n_j$, where n_j is the number of images in the j^{th} image-set: $X_j = \{x_j^i\}_{i=1}^{n_j} \in \mathcal{R}^{d\times n_j}$. The individual samples of a set are described with d dimensional feature representations and denoted as $x_j^i \in \mathcal{R}^d$. The proposed approach is generic and any feature representation may be used, for example, the hand crafted features such as LPB and HoG or automatically learned features such as deep CNN features.

The mean of an image-set X_j is used to capture the first order statistics $\mu_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_j^i$, and the covariance matrix Σ_j is used to capture the second order set statistics

$$\Sigma_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}[x_j^i - \mu_j][x_j^i - \mu_j]^\top. \tag{6}$$

If the number of images is less than the dimensionality of the feature vector, $n_j \leq d$, then the covariance matrix $\Sigma_j \in \mathcal{R}^{d\times d}$ will become rank deficient and will be SPD. In this work, we propose to decompose Σ_j using Cholesky decomposition. Cholesky decomposition yields a unique solution only if the input matrix is positive definite. To ensure a unique decomposition, all eigenvalues of the input matrix must be positive, which we obtain by introducing a regularization term

$$\widehat{\Sigma}_j = \Sigma_j + \frac{\sum_{i=1}^{d}\lambda_j}{\tau} I, \tag{7}$$

where λ_j are the eigenvalues of Σ_j, I is an identity matrix of the same size as that of Σ_j, and $\tau > 1$ is a positive constant. To simplify the analysis, in all experiments, we use a fixed value of $\tau = 1000.00$. Since SVD is computationally expensive and efficiency is an important criterion of our approach, we avoid SVD and add a fraction of the sum of leading diagonal of the covariance matrix. The value of $\sum\lambda_j$ is computed as $\sum\lambda_j = \text{trace}(\Sigma_j)$.

Applying Cholesky decomposition on $\widehat{\Sigma}_j$ we get $\widehat{\Sigma}_j = \Lambda_j \times \Lambda_j^\top$ where Λ_j is a lower triangular matrix with all diagonal entries ≥ 0. We obtain two types of multi-order statistical descriptors from Λ_j and μ_j. The first type of descriptor is computed by adding μ_j to each column of Λ_j

$$\widehat{\Lambda}_j = \Lambda_j + \mu_j 1^{1\times d}, \tag{8}$$

where $\mu_j 1^{1\times d}$ results in a matrix of size $d \times d$ with repeating columns as μ_j. The feature vector $\widehat{f}_j$ is obtained by applying a function $\psi()$ on $\widehat{\Lambda}_j$ to pick the upper triangular values

$$\mathcal{F}_j = \psi(\widehat{\Lambda}_j) = \begin{cases} \widehat{\Lambda}_j(x,y) & \text{if } x \geq y \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

We rearrange $\mathcal{F}_j$ as a vector $\mathcal{F}_j \in \mathcal{R}^{\frac{d(d+1)}{2}}$. For the second type of descriptor we first arrange non-zero entries of Λ_j in to a vector u_j and then concatenate μ_j with u_j:

$$\mathcal{V}_j = [u_j^\top\ \mu_j^\top]^\top. \tag{10}$$

The dimensionality of $\mathcal{V}_j$ is $\frac{1}{2}d(d+3)$. Both $\mathcal{F}_j$ and $\mathcal{V}_j$ are global descriptors therefore, the distance between two image-sets X_i and X_j can be efficiently computed by computing the distance between the corresponding descriptors $\mathcal{F}_i$ and $\mathcal{F}_j$ in the Euclidean space. We have empirically observed that the proposed descriptors are more discriminative than the existing descriptors such as convex hull or affine hull based image-set representations [9], [10] or manifold set representations [8], [31].

C. MLDA FOR RANK DEFICIENT MATRICES

The multi-order descriptors obtained in the last section have relatively large dimensionality. We further reduce the dimensionality of these descriptors for computational efficiency while maintaining the discrimination by using Multiple Linear Discriminant Analysis (MLDA).

Let $G_p = \{\mathcal{F}_{kj}\}_{k=1,j=1}^{m_j,c} \in \mathcal{R}^{\beta\times g}$ be the compact gallery representation consisting of one feature vector for each image-set. Here $\beta = d(d+1)/2$ for $\mathcal{F}_j$ and $\beta = d(d+3)/2$ for $\mathcal{V}_j$, c is the number of object categories contained in the gallery, $m_j \geq 1$ is the number of image-sets in each category, $g = \sum_{j=1}^{c} m_j$, and $\mathcal{F}_{kj}$ is the k^{th} feature vector of the j^{th} class. We intend to reduce the dimensionality of the $\mathcal{F}_{ij}$ or $\mathcal{V}_{ij}$ from β to $c-1$, by using MLDA.

MLDA may be considered as a generalization of the Fisher Linear Discriminant Analysis (FLDA) which compute $c-1$ discriminative directions for classification over c classes. Therefore, our proposed features $\mathcal{F}_{kj}$ are projected from the β dimensional input space to a $c-1$ dimensional discriminative feature space: $\widehat{\mathcal{F}}_{kj} = W^\top\mathcal{F}_{kj}$. The transformed gallery can thus be denoted by: $\widehat{G}_p = W^\top G_p \in \mathcal{R}^{(c-1)\times g}$, where W is the learned projection matrix in $\beta\times(c-1)$, learned by solving an optimization problem for minimizing the intra-class scatter and maximizing the inter-class scatter of $\widehat{G}_p$.

Let S_j denotes the intra-class scatter for the j^{th} category

$$S_j = \sum_{k=1}^{m_j}(\mathcal{F}_{kj} - \boldsymbol{\mu}_j)(\mathcal{F}_{kj} - \boldsymbol{\mu}_j)^\top \tag{11}$$

where $\boldsymbol{\mu}_j$ is the mean of the j^{th} category. The rank of the outer product $(\mathcal{F}_{kj} - \boldsymbol{\mu}_j)(\mathcal{F}_{kj} - \boldsymbol{\mu}_j)^\top$ is one, therefore the rank of S_j will be upper bounded by m_j: $\mathbb{R}(S_j) \leq m_j$. The total within-class scatter is computed by the summation of the class-scatter matrices $S_w = \sum_{j=1}^{c} S_j$ and the rank of $S_w \in \mathcal{R}^{\beta\times\beta}$ is upper bounded by the sum of ranks of the individual class-scatter matrices and therefore by the number of image-sets in the gallery, which shows that S_w is essentially a rank deficient

symmetric matrix. The inter-class scatter matrix is given by

$$S_b = \sum_{j=1}^{c} m_j[\boldsymbol{\mu}_j - \boldsymbol{\mu}][\boldsymbol{\mu}_j - \boldsymbol{\mu}]^\top \tag{12}$$

where $\boldsymbol{\mu}$ is the overall average of the gallery feature vectors and $\boldsymbol{\mu}_j$ is the mean of each class. The intra-class scatter matrix S_w is obtained by summation over c rank-one matrices, therefore it can have maximum c non-zero eigenvalues.

The inter-class scatter of the transformed feature vectors is $\widehat{S}_b = W^\top S_b W$ and intra-class scatter is $\widehat{S}_w = W^\top S_w W$. Conventionally, W is computed such that the ratio $\widehat{S}_b/\widehat{S}_w$ is maximized. Using the fact that the determinant of a matrix is the product of its eigenvalues which also represents its scatter therefore, the ratio of the determinants of the corresponding matrices can be maximized

$$W^* \equiv \arg\max_W \frac{|W^\top S_b W|}{|W^\top S_w W|}. \tag{13}$$

W^* is the set of generalized eigenvectors of S_b and S_w corresponding to the $c-1$ dominant eigenvalues:

$$S_b W = \Lambda S_w W, \tag{14}$$

where Λ contains eigenvalues on the leading diagonal while the rest of the entries are zero. For a non-singular S_w, W may be computed as the eigenvectors of $S_w^{-1}S_b$. Since S_w is rank deficient therefore it is not invert-able and the traditional solution cannot be directly applied. If each image-set results in an independent feature vector, the null space of S_w will have $(\beta - g) \times \beta$ dimensions. Therefore, minimization process of intra-class scatter, $W^\top S_w W$ can find a W within the null space of S_w, such that $S_w W = 0$ and $|W^\top S_w W| = 0$, yielding $|\widehat{S}_b|/|\widehat{S}_w| \rightarrow \infty$, without properly maximizing $|\widehat{S}_b|$. This degenerated case of LDA needs to be properly handled. In other similar problems, regularization is applied on the rank deficient matrix to make it positive definite. We consider it an opportunity to reduce dimensionality of the feature space. Therefore, in the current work we propose to reduce the dimensionality of features by using PCA such that S_w becomes full rank [38].

The rows of the gallery matrix G_p are significantly larger than the columns: $\beta > g$. The row-rank and column-rank of a matrix are always the same therefore, the number of linearly independent rows in G_p are also $\leq g$. To remove row redundancy, we propose to use PCA. To save computation time, we perform the PCA on G_p using the Eigenfaces algorithm [39]. We learn PCA basis such that the total scatter $\widehat{S}_w + \widehat{S}_b$ is maximized

$$\Psi^* \equiv \arg\max_\Psi |\Psi^\top (S_w + S_b)\Psi| \tag{15}$$

or $(S_w + S_b)\Psi = \Lambda\Psi$, which shows that Ψ contains eigenvectors of $S_w + S_b$.

Both scatter matrices S_w and S_b are transformed by Ψ as given by $\Psi^\top S_w \Psi$ and $\Psi^\top S_b \Psi$. The size of both projected scatter matrices is $(g-c) \times (g-c)$, and the rank of $\Psi^\top S_w \Psi$ is $g - c$. Transformation matrix Φ is computed using these reduced dimensionality scatter matrices

$$\Phi^* \equiv \arg\max_\Phi \frac{|\Phi^\top \Psi^\top S_b \Psi \Phi|}{|\Phi^\top \Psi^\top S_w \Psi \Phi|}. \tag{16}$$

Solution is given as Generalized Eigenvalue problem

$$\Psi^\top S_b \Psi \Phi = \Lambda \Psi^\top S_w \Psi \Phi. \tag{17}$$

For a non-singular $\Psi^\top S_w \Psi$, Φ is computed as eigen-vectors of $(\Psi^\top S_w \Psi)^{-1}(\Psi^\top S_b \Psi)$. To compute a compact as well as discriminative gallery representation, we transform the gallery matrix G_p with $W = \Psi\Phi$: $\widehat{G}_p = (\Psi\Phi)^\top G_p$, $\widehat{G}_p \in \mathcal{R}^{c-1 \times g}$ is then used for test image set classification.

The test image-set is also projected on $\Psi\Phi$ as follows: $\mathcal{F}_{pt} = (\Psi\Phi)^\top \mathcal{F}_t$, and the distance of $\mathcal{F}_{pt}$ is independently computed from each feature vector in the gallery $\widehat{G}_p$

$$l_p \equiv \min_{1 \leq j \leq c} \Big(\min_{1 \leq k \leq m_j} (||\widehat{\mathcal{F}}_{kj} - \mathcal{F}_{pt}||_2) \Big), \tag{18}$$

where l_p is the estimated label of the probe image set and $||\cdot||_2$ is the ℓ_2 norm.

D. KERNEL LINEAR DISCRIMINANT ANALYSIS

In this section, we propose Kernel Linear Discriminant Analysis (KLDA) for dimensionality reduction [40] as well as classification accuracy improvement. We want to reduce the dimensionality of the features $\mathcal{F}_{kj}$ (or $\mathcal{V}_{kj}$) from β to $c-1$ using KLDA. We consider ϕ as a non-linear function that maps the feature vectors $\mathcal{F}_{kj} \in \mathcal{R}^\beta$ to a high dimensional space

$$\phi : \mathcal{R}^\tau \mapsto \mathcal{H}. \tag{19}$$

Assuming the mapped data is centred [41], let G_h be the gallery in the higher dimensional feature space $G_h = \phi(G_p) = \{\phi(\mathcal{F}_{kj})\}_{k=1,j=1}^{m_j,c}$. The kernel matrix $K \in \mathcal{R}^{g \times g}$ is defined using dot products between samples in the high dimensional feature space $K = {G_h}^\top G_h$. However, in practice, there is often no need to explicitly define the nonlinear mapping ϕ and the kernel matrix K can be computed from a kernel function in the input space.

Once K is known, KLDA intends to maximize the following objective function

$$\alpha_{opt} = \arg\max \frac{\alpha^\top K \mathcal{W} K \alpha}{\alpha^\top K K \alpha} \tag{20}$$

where $\alpha = [\alpha_1, \ldots, \alpha_g]^\top$, $\mathcal{W} \in \mathcal{R}^{g \times g}$ is a matrix having block diagonal structure: $\mathcal{W} = diag\{\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_c\}$, and $\mathcal{W}_j \in \mathcal{R}^{m_j \times m_j}$ has all elements equal to $1/m_j$. The most significant eigenvector of the following equation yields the optimal α:

$$(KK + \epsilon I)^{-1}(K\mathcal{W}K)\alpha = \lambda\alpha, \tag{21}$$

If the kernel matrix K is rank deficient, it is regularized by adding a small ϵ to the leading diagonal entries yielding KK an invert-able matrix. By selecting the $(c-1)$ most significant eigenvectors, a transformation matrix is obtained

FIGURE 2. (a) HONDA/UCSD: Image samples from different image sets are displayed in rows. (b) CMU MoBo: Two image sets from each class are displayed. (c) Youtube dataset: Two example image sets from each class are displayed.

$\widehat{\alpha} = [\alpha_1, \ldots, \alpha_{c-1}]$. Let $\mathcal{F}_t \in \mathcal{R}^{\beta}$ be the test feature vector in the input space whose mapping in the feature space is $\phi(\mathcal{F}_t)$. The KLDA feature $\mathcal{Y}_t \in \mathcal{R}^{c-1}$ in the discriminant subspace is given by

$$\mathcal{Y}_t = \widehat{\alpha}^{\top} G_h^{\top} \mathcal{F}_t. \tag{22}$$

KLDA is non-linear in the input space because of the non-linear mapping ϕ between the input and the feature space. Non-linear mapping can increase the discrimination ability of a classifier, according to Cover's theorem on the separability of patterns [42]. Due to the possibly very high computational cost, the nonlinear mapping function ϕ is never implemented explicitly [41], [43], [44]. For this purpose, kernel functions are applied in the input space to achieve the same effect of the computationally expensive nonlinear mapping. The kernel functions allow to compute the scalar product implicitly in $\mathcal{H}$, without explicitly using or even knowing the mapping ϕ [44]. However, for a given function to be a kernel function, it must satisfy the Mercers condition [41], [44]. KLDA takes advantage of this kernel trick and computes the inner products by means of a kernel function. The polynomial kernel is one of the widely-used kernel function that fulfils the Mercer's condition and is given by [44]

$$k(\mathcal{F}_{kj}, \mathcal{F}_{k'j'}) = (\mathcal{F}_{kj} \cdot \mathcal{F}_{k'j'})^{\beta}. \tag{23}$$

where β is the order of polynomial.

III. EXPERIMENTAL EVALUATION

The two types of multi-order statistical descriptors and two classifiers make four different combinations $\{\mathcal{F}_j + MLDA, \mathcal{F}_j + KLDA, \mathcal{V}_j + MLDA, \mathcal{V}_j + KLDA\}$. We perform extensive evaluation experiments to assess the performance of these descriptors. We use six benchmark datasets for face recognition, object categorization, periocular biometric recognition and hand gesture recognition in our experiments. These datasets include Honda/UCSD [45], CMU Mobo [46] and Youtube celebrities' datasets for holistic face identification and are widely used in the literature to evaluate the performance of image set classification techniques. We use the MBGC v2 NIR video dataset for periocular biometric recognition. For the object categorization experiments we use the ETH-80 [47] dataset. Hand gesture recognition experiments are performed on the Cambridge dataset [48]. We compare the multi-order statistical descriptors with 10 existing object classification algorithms for classification accuracy and execution time comparisons. Our experimental results verify the superiority of the proposed descriptors over the existing algorithms.

A. DESCRIPTION OF THE DATSETS

The Honda/UCSD dataset [45] contains 20 different subjects distributed over 59 videos. We extracted the faces in these video sequences using Viola and Jones method [49]. We re-sized the gray-scale images to 20 × 20 pixels and applied histogram equalization to reduce illumination variations. Fig. 2-a shows cropped and re-sized face images from Honda dataset.

There are 96 whole body video sequences of 24 different persons (classes) in the CMU MoBo dataset. For our experiments, we used [49] for automatically detecting and cropping the facial regions. The video sequences on which the face detector failed were manually processed for cropping the faces. We filter the gray-scale face images with a circular (8,1) neighborhood LBP filter [50] and re-size the LBP coded images to 20 × 20. Note that this is different from [9] where they used the histograms of the LBP features. The feature dimension as a result of the LBP coding is smaller $d = 400$ and achieves better classification accuracy in our experiments. Some face images from the CMU-MoBo dataset are shown in Fig. 2-b.

The Youtube Celebrities dataset [51] is one of the most challenging benchmarks for evaluating image set classification algorithms. This dataset is comprised of 1910 videos of 47 different media celebrities such as film actors, actresses and politicians etc. These videos were collected from the Youtube website (Fig. 2-c). Each video sequence has different numbers of image frames ranging from 8 to 400. Individual frames are low in resolution and has high compression ratios due to which the quality is degraded and most image frames are noisy. Moreover, large facial pose variations,

FIGURE 3. Examples of periocular image sets belonging to two different subject lasses (left and right) of the MBGC NIR v2 database.

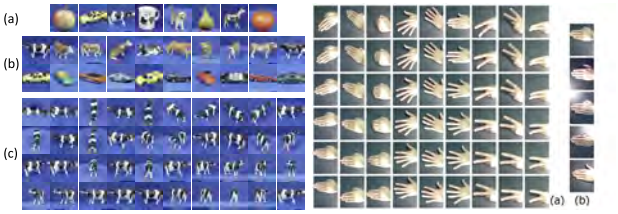

FIGURE 4. (Left) ETH-80 object categories dataset. (a) Object categories available in the dataset. (b) 10 different types of objects to represent each category. (c) Individual image samples of a set from the cow category. (Right) Cambridge Hand Gesture database. (a) Sample image frames from nine gesture classes. Each column represents a different gesture class. (b) Different illumination settings used to capture the database.

difficult illumination conditions and varying facial expressions are present in the videos. For our experiments, we automatically extracted face images in the video sequences by applying [49] and re-sized them to 20×20 pixels. We then computed the LBP histogram features of the gray-scale face images using a cell size of 5. As a result, feature vectors of $d = 928$ dimensions are obtained for each face image. For LBP histogram feature extraction, we used the VLFeat library [52].

We use the near-infrared (NIR) video sequences provided by the MBGC portal challenge v2 [53] for periocular biometric recognition. This dataset was captured while subjects were walking through a portal towards the camera. NIR illumination was used to capture high resolution iris videos. Most frames in the videos exhibit out-of-focus motion blur, large variations in illumination, sensor noise, and specular reflections. We automatically extracted the periocular region consisting of two eyes and the nose bridge using [49]. The detected regions are scale and rotation normalized based on the automatically detected pupil centers. Separate left and right regions are extracted from the normalized periocular images. This experimental setting simulates the real-world situations when only one or both eyes may be visible while the rest of the face is occluded. The right periocular region is mirrored to the left and the resulting images are normalized to unit magnitude to reduce the effect of illumination variations. Figure 3 shows some example sample images of two different subjects from MBGC NIR dataset.

The ETH-80 dataset [47] consists of image sequences of 8 different object categories (Fig. 4). Each category consists of 10 different object types belonging to the same general class. Each object is imaged from 41 different views and thus creating an 41 sample image-set. We use re-sized images of size 20×20 in our experiments. This is a challenging dataset for image set classification due to few images in each set, significant within-class appearance differences and large view angle variations in individual image samples.

TABLE 1. Benchmark datasets used in our experiments. Min, Max and Avg represent the minimum, maximum and average image samples available in each set, respectively.

Dataset	Classes	Sets/class	Min	Max	Avg
Honda/UCSD	20	1-5	13	782	267
CMU-Mobo	24	4	68	370	307
MGBC	114	1-12	6	48	18
Youtube Celeb	47	17-108	7	350	155
ETH-80	8	10	41	41	41
Cambridge	9	100	37	119	71

The Cambridge Hand Gesture dataset [48] (Fig. 4) consists of nine hand gesture classes distributed over 900 video sequences. These nine classes are created from a combination of 3 basic hand shapes and motions patterns. Each class has 100 video sequences captured in five different illumination settings, 10 arbitrary motion patterns and were executed by 2 persons. A fixed camera was used to capture each hand gesture. Significant sample and sequence wise within-class variations are present in the dataset. By following the standard experimental protocol defined by [54], the 100 video sequences available for each gesture class are divided based on the 5 illumination settings (Set-1, Set-2, Set-3, Set-4 and Set-5). Set-5 is selected for training while the rest are used for testing. We used gray scale frames of size 20×20 without any pre-processing for our experiments. Table (1) shows the dataset details used in our experiments.

B. EXPERIMENTAL SETUP

We compare the proposed algorithms with 11 object classification techniques including Canonical Correlation Analysis (DCC) [29], Covariance Discriminant Learning (CDL) [11], Manifold to Manifold Distance (MMD) [31], Regularized Nearest Points (RNP) [27], Manifold Discriminant Analysis (MDA) [8], Mean Sequence Sparse Representation Classification (MSSRC) [55], the Linear Affine Hull-based Image Set Distance (AHISD) [9], Sparse Approximated Nearest Points (SANP) [10], the Convex Hull-based Image Set Distance (CHISD) [9], and Set to Set Distance Metric Learning (SSDML) [56]. Standard implementations provided by the original authors are used in our experiments. However, Hu's [10] implementation of MDA is used, while CDL is self-implemented. We use the standard experimental protocol defined previously by [8]–[11], and [31] to conduct our experiments.

We carefully choose the hyper parameters of each technique involved in our study. For DCC, a 10-dimensional subspace is used to represent image sets. Similarly, 10 maximum canonical correlations are used for discriminative learning. For MMD and MDA, we follow the recommendations of [31] and [8] to configure the hyper parameters. The ratio of Euclidean and geodesic distances is optimized for each dataset. We search different values in the range {1.0-5.0} and

TABLE 2. Comparison of the average and standard deviation (%) of image set classification accuracy achieved by different algorithms in 10-fold experiments on the five datasets and 5 fold experiments on Youtube dataset.

	Honda	MoBo	ETH-80	MBGC	Youtube	Cambridge
DCC [29]	94.67±1.32	93.61±1.76	73.33±4.03	76.85±0.51	66.75±3.47	69.25±2.67
MMD [31]	94.87±1.16	93.19±1.66	69.72±4.01	64.35±2.61	65.12±4.36	32.17±2.71
MDA [8]	97.44±0.91	97.06±1.02	45.53±4.56	91.01±1.59	68.12±4.85	26.85±4.36
CDL [11]	100.0±0.00	95.83±2.07	75.00±4.26	75.65±1.49	68.96±5.29	78.47±3.14
AHISD [9]	89.74±1.85	97.36±0.79	51.52±5.92	87.38±2.05	71.92±3.55	23.11±3.91
CHISD [9]	92.31±2.12	96.41±0.97	51.67±4.11	88.04±1.52	73.17±3.29	25.31±2.51
SANP [10]	93.08±3.43	96.94±0.63	49.17±3.83	88.33±1.39	74.04±3.48	25.45±2.23
MSSRC [55]	96.75±2.65	97.05 ± 0.88	67.50±3.07	89.49±2.88	74.24±3.21	28.85±2.11
SSDML [56]	89.41±3.64	85.75 ± 1.82	73.20±2.12	70.52±1.87	70.81±3.42	32.24±2.85
RNP [27]	95.95±2.16	96.11 ± 1.43	50.21±3.24	88.50±2.17	74.02±3.68	22.03±2.91
$\mathcal{F}_j$+MLDA	**100.0±0.0**	97.36±0.79	72.91±2.38	92.32±1.35	76.17±3.34	84.38±3.17
$\mathcal{V}_j$+MLDA	**100.0±0.0**	97.50±0.80	79.58±3.70	81.90±0.74	77.14±3.23	88.51±1.98
$\mathcal{F}_j$+KLDA	**100.0±0.0**	97.50±0.80	73.19±3.98	**93.33±0.62**	76.88±3.74	87.72±1.09
$\mathcal{V}_j$+KLDA	**100.0±0.0**	**97.64±0.67**	**80.00±4.15**	83.33±0.60	**77.19±2.98**	**89.64±1.73**

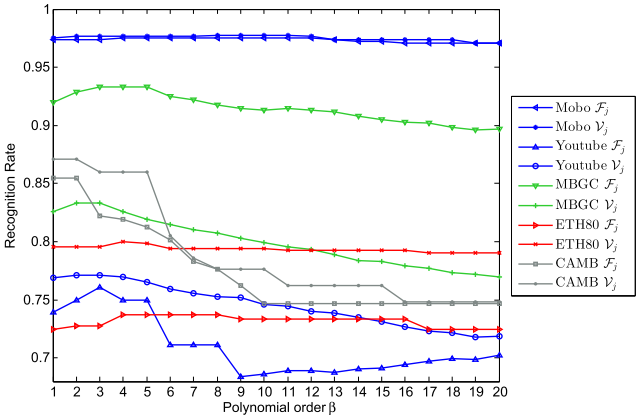

FIGURE 5. Accuracy of KLDA versus the order of polynomial β in (23). For the Honda dataset accuracy remained 100% for all values of β. For the Mobo, Youtube, MBGC, ETH and Cambridge datasets, the highest accuracy was achieved for $\beta = \{2, 3, 3, 4, 2\}$ respectively.

report the best results. The top most canonical correlation is used to calculate the MMD. A search space of {10, 12, 15, 18} is used to find the best number of connected nearest neighbours for geodesic distance in MDA and MMD. Similarly, a search space of {80%, 85%, 90%, 95%, 99%} is used to select the best value for the number of PCA basis used to represent each image set in AHISD, CHISD and SANP. Parameter C is set to 100 in the SVM optimization framework of CHISD. For RNP [27], 90% PCA energy is preserved and same weight parameters are used as in [27]. MSSRC [55] and SSDML [56] are parameters free.

For Honda, MoBo, MBGC and ETH-80 data sets, we used one image set from each class to construct the gallery while the remaining are used for testing. We performed 10-fold experiments, each time randomly selecting gallery and test set combinations. For Youtube dataset, we perform experiments based on the standard experimental protocol of [10]. Specifically, five-fold cross validation experiments are designed in which the complete dataset is divided equally into five disjoint parts. Each fold consists of nine image sets for each class. In each fold, gallery is constructed by randomly selecting three image sets per class while the remaining six image sets are used as test sets. For Cambridge hand gesture dataset, experiments are performed based on the protocol defined by [54]. Specifically, five sets (Set-1, Set-2, Set-3, Set-4 and Set-5) are created based on the five illumination settings. Set-5 of each class is used for training. The training image sets are further partitioned randomly into gallery sets and validation sets. Specifically, 10 image sets are chosen for the gallery while the other 10 image sets are set aside for validation. Experiments are repeated 10-folds with different combinations of gallery and validation sets in each fold.

The learning process of MLDA and KLDA require at least two samples from every class. Therefore, for the classes having only a single image set available in the gallery, we construct two disjoint image sets from the single one by randomly partitioning it. In our experiments, we preserve 100% energy of the basis, because all discarded basis had zero singular values. In KLDA based classification, we use the polynomial kernel (23) to report the results. We perform analysis of KLDA accuracy for the appropriate choice of the polynomial order β. Fig. 5 shows accuracy variations as the order is changed from 1 to 20. For the Mobo, Youtube, MBGC, ETH and Cambridge datasets, the highest accuracy was achieved by setting $\beta = \{2, 3, 3, 4, 2\}$ respectively. The code to compute the proposed descriptors will soon be made publicly available.

C. IMAGE SET CLASSIFICATION RESULTS

Table 2 summarizes the results of our image set classification experiments using the six benchmark datasets. In the case of Honda/UCSD dataset, all four combinations of our proposed descriptors achieved 100% accuracy and outperformed the comparative methods. Please note that the accuracy of SANP, AHISD, and CHISD is lower compared to that reported in [9] and [10]. This is because we evaluate these algorithms in 10-fold experiments whereas these were evaluated in a single fold in [9] and [10] where the first image set of each subject was chosen for gallery and the rest were used as probes. Also, we use 20 × 20 images in our experiments whereas the image size was 40 × 40 in [9] and [10].

The CMU-Mobo dataset has relatively smaller with-in class facial pose changes and the effect of illumination has been normalized using LBP filtering. Thus, the mean of each image set is more discriminative compared to the other datasets. The proposed descriptors $\mathcal{F}_j$ and $\mathcal{V}_j$ performed better than CDL which is based on 2-nd order statistic (Table 2). AHISD, MDA and MSSRC also perform good on this dataset. These experiments show that the multi-order descriptor performs better than single order descriptors used in CDL, AHISD, MDA and MSSRC. Within the four combinations of proposed descriptors, the performance of $\mathcal{V}_j$ + KLDA is best than the others.

On the Youtube celebrities dataset, all four combinations of the proposed descriptors outperformed the existing methods (Table 2). The image sets in this dataset are relatively more noisy and their structure cannot be perfectly estimated. Therefore, the structure based algorithms (DCC, CDL) perform poor compared to sample based algorithms (AHISD, CHISD, SANP, MSSRC, RNP). In contrast, the proposed descriptors combine both the sample as well as the structural properties of the image sets and are therefore more accurate than the existing methods. Our use of the LBP histogram features increases the discrimination. Therefore, most of the existing algorithms achieve relatively higher accuracy than previously reported on this dataset [10], [11].

In the MBGC NIR database, there are fewer frames (4-48 images) per image set therefore, the structure based algorithms perform poor on this dataset (Table 2). CDL shows less accuracy because the covariance matrix estimate from a small number of frames is poor and due to ignoring the mean descriptor. Similarly, DCC and MMD also fail to accurately model the image set in this dataset. The mean of each image set is informative in MBGC NIR dataset due to the better alignment of periocular regions. For MBGC NIR database, the fusion of the mean with the second order statistics using summation ($\mathcal{F}_j$) outperforms all the other algorithms. This shows that when the covariance estimate is poor the addition of mean ($\mathcal{F}_j$) is better than the concatenation ($\mathcal{V}_j$).

We use only one image set per class in the gallery for training over ETH-80 dataset. This render the classification problem more difficult than using five image sets per gallery as was the case in DCC [29], MDA [8] and CDL [11]. On this dataset, the sample based algorithms perform poor due to the large intra class pose variations and significant intra-class object appearance differences. The proposed multi-order statistical descriptor $\mathcal{V}_j$ outperforms the existing algorithms when used with both MLDA and KLDA (Table 2).

The results on the Cambridge Hand Gestures dataset show that the proposed descriptors are generic. We apply exactly the same algorithm as are applied for the image set classification. Table 2 shows that the proposed multi-order statistical descriptors ($\mathcal{F}_j$, $\mathcal{V}_j$) outperformed the other image set representation methods for the task of hand gesture recognition on this dataset. While other methods for gesture recognition such as [57] and [58] have shown greater accuracy on the Cambridge Hand Gesture dataset, these algorithms are not tested for the generic task of image set based classification. The sample based algorithms (AHISD, SANP, CHISD, MSSRC, RNP) perform poor on this dataset. This is because the sample based algorithms use the location of individual sample to define their models which cannot express the hand gesture dynamics robustly in their representations. The proposed descriptors fuse both the structure and the sample measures of the image set, can better model the hand gesture variations. DCC and CDL also achieved more accuracy than the sample based algorithms on this dataset which shows that structure based algorithms have an edge over the sample based algorithms for the task of gesture recognition.

D. ROBUSTNESS EXPERIMENTS

We used the Honda/UCSD dataset for robustness experiments. We first evaluated the proposed algorithm for its robustness to the number of samples available in each image set for modeling. We randomly selected {10, 20, 30, 40, 50, 60, 70, 80, 90, 100} samples to form a set. The average recognition rates obtained in these experiments are shown in Fig. 6-a. The accuracy of the proposed algorithms is relatively lower when 10 images per set were used. However, as the number of images used to construct a set increases, the accuracy of the proposed algorithms increases dramatically and $\mathcal{F}_j$ + *MLDA* achieves 99.79% recognition rates at 40 images per set. All other algorithms exhibited very different behaviors with the increase in the number of images per set. SANP and CHISD obtained their maximum recognition rates of 96.92 % and 97.69% respectively at 20 images per set and further increase in the number of images per set was mostly unfavorable for these two algorithms. The accuracy of MDA linearly increased until 60 images per set however, the accuracy dropped at 70 images followed by a liner increase. The accuracy of DCC increased in a piecewise liner fashion. The accuracy of MMD increased with a big jump from 10 to 20 images per set however, it quickly reached a saturation value at 50 followed by a decreasing trend for 60 and 70 images per set. A second increasing trend followed for 80 and 90 images per set reaching a saturation level of 92.56 % at 100 images per set. The maximum gain in accuracy for the proposed descriptors was till 50 images per set and a saturation level of 100% accuracy was reached at 60 images for all the variants of the proposed descriptors. This shows that 60 random frames per set are optimal for capturing the first and the second order statistics in this dataset. Moreover, the accuracy of the proposed algorithms monotonically increases with the increase in the number of images used to form a set and also have no negative effects of addition of more samples.

In the next experiment, we evaluated the robustness of the proposed descriptors to the presence of noise and compared it to other algorithms in a setup similar to [9]. Using the Honda dataset, we constructed a clean gallery and test image sets each containing 100 randomly selected images. This is to ensure that the percentage of noise added to each set should be

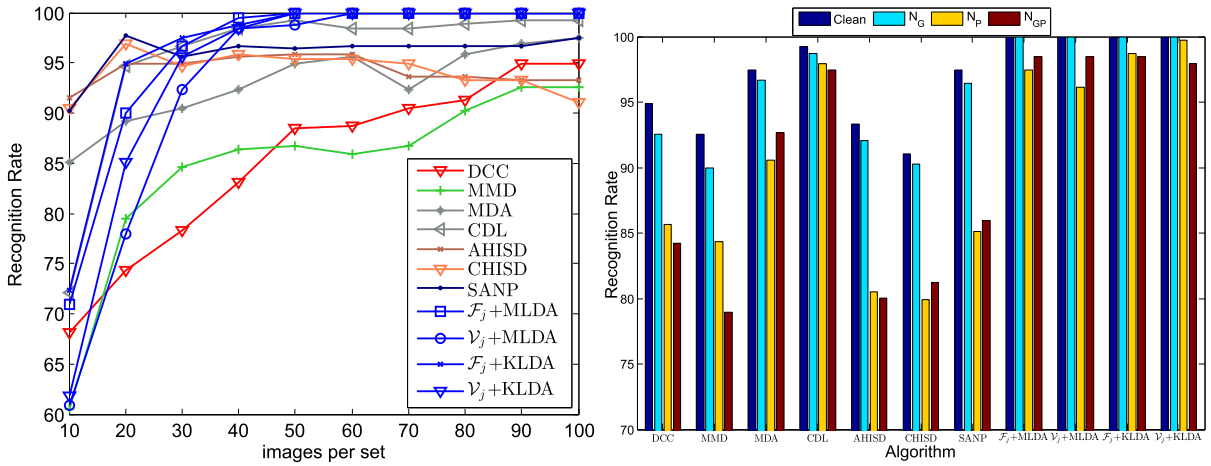

FIGURE 6. (a) Robustness to the number of image samples in the sets. (b) Robustness to noise in the image sets of gallery, probe and both. (Recognition rates (classification accuracy) are average of 10 fold experiments on Honda/UCSD dataset.)

the same. Then, we construct noisy image sets by randomly choosing one image from each image set and adding it to the image sets of the other classes. This is done for both gallery and test image sets to generate three scenarios. The original clean image set data and the three noise levels induced cases were denoted by N_c (clean), N_G (noisy gallery only), N_P (Noisy probes only) and N_{G+P} (noisy gallery and probe). The results in Fig. 6 shows that the proposed descriptors exhibited robustness to noise better than the other algorithms. As expected, the structure based techniques are more robust to noise compared to the sample based techniques (AHISD, SANP, CHISD). This is because the holistic model of set structure has a smoothing effect which reduces the influence of individual noisy samples. In contrast, sample based algorithms usually generate interpolated samples from the original samples. This can lead to in-accurate representation of the set when noisy samples are included in the approximation process.

We performed two more experiments in which we added outliers only in the test sets. The presence of outliers in only test sets is more challenging but realistic scenario because usually the training sets are chosen such that they do not contain outliers. We setup the first experiment such that $(g-1)n_r$ samples are added to each test set, where g is the gallery size and n_r is the number of randomly selected samples from the other gallery sets. By varying n_r from 1 to 3 we added 19, 38 and 57 outliers to each probe set. Table 3 shows a comparison of accuracy of different algorithms for these three challenging cases. The drop in the recognition rate of our proposed descriptors is significantly lower compared to the others. For example, in the case of the proposed $\mathcal{V}_j$+*KLDA* algorithm, the drop in the recognition rate when $n_r = 3$ is 0.5% which is significantly less than the 5.38% drop of CDL and 1.92% drop of SANP. This experiment demonstrated the robustness of the proposed descriptor $\mathcal{V}_j + KLDA$ to noise in the image sets.

TABLE 3. Comparison of the average accuracy of different algorithms in the presence of outliers.

Algorithm	$n_r = 1$	$n_r = 2$	$n_r = 3$
CDL [11]	98.72 ± 1.28	96.92 ± 2.65	94.62 ± 1.89
MDA [8]	97.44 ± 0.91	96.73 ± 1.14	95.73 ± 2.87
DCC [29]	93.59 ± 2.78	92.93 ± 2.25	92.31 ± 2.09
AHISD [9]	88.21 ± 1.09	87.31 ± 2.42	87.03 ± 1.35
MMD [31]	93.83 ± 1.16	93.04 ± 2.44	89.74 ± 4.05
CHISD [9]	92.11 ± 1.89	91.81 ± 2.42	91.03 ± 1.35
SANP [10]	92.82 ± 1.32	91.54 ± 3.83	91.16 ± 2.84
$\mathcal{F}_j$+MLDA	100.00 ± 0.00	100.00 ± 0.00	94.87 ± 1.73
$\mathcal{V}_j$+MLDA	100.00 ± 0.00	100.00 ± 0.00	98.12 ± 1.79
$\mathcal{F}_j$+KLDA	100.00 ± 0.00	98.97 ± 1.03	96.92 ± 2.02
$\mathcal{V}_j$+KLDA	100.00 ± 0.00	100.00 ± 0.00	99.49 ± 0.51

In our second experimental setup, we evaluate the robustness of the proposed descriptors to the presence of strong outliers belonging to one specific class. This is done by adding randomly selected n_r images from a randomly selected gallery set to each probe set. This is a more challenging case because many outliers from the same class can alter the structure of the image set. The value of n_r is varied from 1 to 12. The accuracy of $\mathcal{F}_j$ + KLDA and $\mathcal{V}_j$ + KLDA remained 100% for $n_r \leq 11$ and 99.74% for $n_r = 12$.

E. COMPARISON OF COMPUTATIONAL TIME

Table 4 summarizes the average execution times of all algorithms in our study. The execution time is calculated for classifying one probe image set by matching with the 141 gallery image-sets in the Youtube dataset. The average time of 5-fold experiments is reported for each algorithm. A 3.4GHz Pentium CPU with 8GB RAM and MATLAB implementations are used to conduct these experiments. These comparisons verify that all variants of the proposed descriptors are significantly faster than existing techniques.

For example, the proposed $\mathcal{F}_j$ + *KLDA* is about 327 and 449 times faster than CDL [11] and SANP [10]

TABLE 4. Comparison of the execution times (in seconds) of different algorithms for classifying one probe image set by matching with the 141 gallery image sets in the Youtube dataset.

Algorithm	Training time (sec)	Testing time (sec)
DCC [29]	167.49	8.08
MMD [31]	313.57	78.32
MDA [8]	580.70	201.48
AHISD [9]	N/A	18.10
CHISD [9]	N/A	190.61
SANP [10]	N/A	17.94
CDL [11]	345.88	13.08
MSSRC [55]	N/A	30.82
SSDML [56]	400.01	21.87
RNP [27]	N/A	6.42
$\mathcal{F}_j$+MLDA	11.52	**0.05**
$\mathcal{V}_j$+MLDA	10.63	0.07
$\mathcal{F}_j$+KLDA	5.28	**0.04**
$\mathcal{V}_j$+KLDA	8.21	0.06

respectively. Our use of LBP histogram features increases the discrimination but makes the feature dimension d very high ($d = 928$). Therefore, the existing algorithms suffer from computational complexity as well as space complexity. However, even for such high dimensional features, all the variants of the proposed descriptors are significantly faster. This shows that the proposed descriptors have better scalability for high dimensional and large datasets. Note that SANP is faster than our previous results [12]. This is due to unit normalizing the feature vectors before input to the SANP optimization algorithm. This helped the accelerated proximal gradient method to converge quickly. We have also significantly optimized the implementation of CDL to achieve faster execution times.

Note that for an $n \times n$ matrix, the computational complexity of a sequential implementation of Cholesky decomposition is $O(n^3)$. Fast algorithms are available (e.g. Matlab chol) for computing the Cholesky decomposition. Therefore, it does not present a computational bottleneck in the proposed algorithm while making it theoretically compliant.

IV. CONCLUSION

In this paper multi-order statistical descriptors are proposed to represent image sets. Dimensionality of the descriptors is reduced using MLDA and KLDA using the polynomial kernels. The proposed descriptors are compared with 11 existing algorithms on six datasets. Experimental results demonstrate that the proposed descriptors are computationally efficient, robust and highly accurate for object classification and face recognition tasks. Experiments also demonstrate that the multi-order descriptors are robust to small number of samples per set and the presence of a large number of outliers in the probe sets as well in the gallery sets. In terms of execution time speedup, the proposed descriptors are 107 times faster than the nearest competitor. Therefore, the proposed descriptors can potentially be used for real time face recognition and object classification in videos.

REFERENCES

[1] J. Yang, L. Luo, J. Qian, Y. Tai, F. Zhang, and Y. Xu, "Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156–171, Jan. 2017.

[2] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2017.270039.

[3] Z. Cui, H. Chang, S. Shan, B. Ma, and X. Chen, "Joint sparse representation for video-based face recognition," *Neurocomputing*, vol. 135, pp. 306–312, Jul. 2014.

[4] S. A. A. Shah, M. Bennamoun, and F. Boussaid, "Iterative deep learning for image set based face and object recognition," *Neurocomputing*, vol. 174, pp. 866–874, Jan. 2016.

[5] M. Uzair, A. Mahmood, A. Mian, and C. McDonald, "Periocular region-based person identification in the visible, infrared and hyperspectral imagery," *Neurocomputing*, vol. 149, pp. 854–867, Feb. 2015.

[6] A. Mahmood, A. Mian, and R. Owens, "Semi-supervised spectral clustering for image set classification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 121–128.

[7] Q.-S. Zeng, J.-H. Lai, and C.-D. Wang, "Multi-local model image set matching based on domain description," *Pattern Recognit.*, vol. 47, no. 2, pp. 694–704, 2014.

[8] R. Wang and X. Chen, "Manifold discriminant analysis," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 429–436.

[9] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2567–2573.

[10] Y. Hu, A. Mian, and R. Owens, "Face recognition using sparse approximated nearest points between image sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1992–2004, Oct. 2012.

[11] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proc. CVPR*, Jun. 2012, pp. 2496–2503.

[12] M. Uzair, A. Mahmood, A. Mian, and C. McDonald, "A compact discriminative representation for efficient image-set classification with application to biometric recognition," in *Proc. Int. Conf. Biometrics*, Jun. 2013, pp. 1–8.

[13] T.-K. Kim, O. Arandjelovic, and R. Cipolla, "Boosted manifold principal angles for image set-based recognition," *Pattern Recognit.*, vol. 40, no. 9, pp. 2475–2484, 2007.

[14] X. Chen, H. Ma, C. Zhu, X. Wang, and Z. Zhao, "Boundary-aware box refinement for object proposal generation," *Neurocomputing*, vol. 219, pp. 323–332, Jan. 2017.

[15] J. Seo and H. Park, "Robust recognition of face with partial variations using local features and statistical learning," *Neurocomputing*, vol. 129, pp. 41–48, Apr. 2014.

[16] C. Peng, X. Gao, N. Wang, and J. Li, "Graphical representation for heterogeneous face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 301–312, Feb. 2017.

[17] M. Uzair and A. Mian, "Regularized least-squares coding with unlabeled dictionary for image-set based face recognition," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DlCTA)*, Nov. 2014, pp. 1–7.

[18] M. Uzair, A. Mahmood, and A. Mian, "Sparse kernel learning for image set classification," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 617–631.

[19] A. Shahroudy, T. T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in RGB+D videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2017.2691321.

[20] D. Xu, R. Song, X. Wu, N. Li, W. Feng, and H. Qian, "Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts," *Neurocomputing*, vol. 143, pp. 144–152, Nov. 2014.

[21] X. Zhao, X. Li, C. Pang, and S. Wang, "Human action recognition based on semi-supervised discriminant analysis with global constraint," *Neurocomputing*, vol. 105, pp. 45–50, Apr. 2013.

[22] C.-C. Jia *et al.*, "Incremental multi-linear discriminant analysis using canonical correlations for action recognition," *Neurocomputing*, vol. 83, pp. 56–63, Apr. 2012.

[23] K. Huang, T. Tan, S. Maybank, R. Chellappa, and J. Aggarval, "Guest editorial introduction to the special issue on large-scale video analytics for enhanced security: Algorithms and systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 4, pp. 589–592, Apr. 2017.

[24] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2018.

[25] G. Shakhnarovich, J. W. Fisher, and T. Darrell, "Face recognition from long-term observations," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 851–868.
[26] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1165–1179, Jun. 2017.
[27] M. Yang, P. Zhu, L. Van Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–7.
[28] A. Mahmood and A. S. Mian. (2014). "Semi-supervised spectral clustering for classification." [Online]. Available: https://arxiv.org/abs/1405.5737
[29] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.
[30] A. Mahmood and A. S. Mian, "Hierarchical sparse spectral clustering for image set classification," in *Proc. BMVC*, 2012, pp. 1–11.
[31] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[32] M. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2705–2712.
[33] T. Wang and P. Shi, "Kernel Grassmannian distances and discriminant analysis for face recognition from image sets," *Pattern Recognit. Lett.*, vol. 30, no. 13, pp. 1161–1165, 2009.
[34] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *Int. J. Comput. Vis.*, vol. 66, no. 1, pp. 41–66, 2006.
[35] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM J. Matrix Anal. Appl.*, vol. 29, no. 1, pp. 328–347, 2007.
[36] W. B. Wu and M. Pourahmadi, "Nonparametric estimation of large covariance matrices of longitudinal data," *Biometrika*, vol. 90, no. 4, pp. 831–844, 2003.
[37] X. Hong, H. Chang, S. Shan, X. Chen, and W. Gao, "Sigma set: A small second order statistical region descriptor," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1802–1809.
[38] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.
[39] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 1991, pp. 586–591.
[40] W. Li, Q. Ruan, and J. Wan, "Dimensionality reduction using graph-embedded probability-based semi-supervised discriminant analysis," *Neurocomputing*, vol. 138, pp. 283–296, Aug. 2014.
[41] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA, USA: MIT Press, 2002.
[42] S. Haykin, *Neural networks: A comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1999.
[43] Z. Sun, J. Li, and C. Sun, "Kernel inverse Fisher discriminant analysis for face recognition," *Neurocomputing*, vol. 134, pp. 46–52, Jun. 2014.
[44] K. R. Müller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.
[45] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proc. CVPR*, vol. 1. 2003, pp. I313–I320.
[46] R. Gross and J. Shi, "The CMU motion of body (MoBo) database," Robot. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-RI-TR-01-18, 2001.
[47] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. II-409–II-415.
[48] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *Proc. CVPR*, Jun. 2007, pp. 1–8.
[49] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[50] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[51] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. CVPR*, Jun. 2008, pp. 1–8.
[52] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.
[53] *Multiple Biometric Grand Challenge (MBGC) Dataset*. Accessed: Nov. 30, 2017. [Online]. Available: http://face.nist.gov/mbgc/
[54] T.-K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1415–1428, Aug. 2009.
[55] E. G. Ortiz, A. Wright, and M. Shah, "Face recognition in movie trailers via mean sequence sparse representation-based classification," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3531–3538.
[56] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: Extend the learning of distance metrics," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2664–2671.
[57] Y. M. Lui, J. R. Beveridge, and M. Kirby, "Action classification on product manifolds," in *Proc. CVPR*, Jun. 2010, pp. 833–839.
[58] Y. M. Lui, "Human gesture recognition on product manifolds," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 3297–3321, 2012.

ARIF MAHMOOD received the master's and Ph.D. degrees (Hons.) in computer science from the Lahore University of Management Sciences in 2003 and 2011, respectively. Before that, he was a Research Assistant Professor with the School of Mathematics and Statistics, and with the College of Computer Science and Software Engineering, The University of the Western Australia. He is currently a Post-Doctoral Researcher with the Department of Computer Science and Engineering, Qatar University. He is keenly interested in exploring the applications of machine learning techniques for the complex network structure characterization. His major research interests are in computer vision and pattern recognition, and more specifically, in data clustering, classification, action and object recognition using image sets, scene background modeling, person segmentation and action recognition in crowds, the computation elimination algorithms for fast template matching, image segmentation, and facial expression mapping.

MUHAMMAD UZAIR received the Ph.D. degree from The University of Western Australia and the M.S. degree from Hanyang University, South Korea. He is currently an Assistant Professor with the Department of Electrical Engineering, COMSATS Institute of Information Technology Wah Campus, Pakistan. His primary research areas are computer vision and machine learning. His research interests include hyperspectral image analysis, spectroscopy, face recognition, domain adaptation, and video-based classification.

SOMAYA AL-MAADEED received the Ph.D. degree in computer science from The University of Nottingham, U.K., in 2004. She started as an Assistant Professor with Qatar University. She has been a Visiting Fellow with Northumbria University, U.K., since 2012. She performed research in several areas, including biometrics, writer identification, image processing, and document analysis. She has published around 40 research papers in peer reviewed conferences and journals. She has received a number of research grants. Her team received the Best Performance Prize in ICDAR 2011 Signature Verification and Music Scores competition.