Real Time Surveillance for Low Resolution and Limited-Data Scenarios: An Image Set Classification Approach

Uzair Nadeem*   Syed Afaq Ali Shah†   Mohammed Bennamoun‡   Roberto Togneri§

Ferdous Sohel¶

Abstract

This paper proposes a novel image set classification technique based on the concept of linear regression. Unlike most other approaches, the proposed technique does not involve any training or feature extraction. The gallery image sets are represented as subspaces in a high dimensional space. Class specific gallery subspaces are used to estimate regression models for each image of the test image set. Images of the test set are then projected on the gallery subspaces. Residuals, calculated using the Euclidean distance between the original and the projected test images, are used as the distance metric. Three different strategies are devised to decide on the final class of the test image set. We performed extensive evaluations of the proposed technique under the challenges of low resolution, noise and less gallery data for the tasks of surveillance, video-based face recognition and object recognition. Experiments show that the proposed technique achieves a better classification accuracy and a faster execution time compared to existing techniques especially under the challenging conditions of low resolution and small gallery and test data.

Keywords

Image Set Classification   Linear Regression Classification   Surveillance   Video-based Face Recognition   Object Recognition

1   Introduction

In the recent years, due to the wide availability of mobile and handy cameras, image set classification has emerged as a promising research area in computer vision. The problem of image set classification is defined as the task of recognition from multiple images (Kim et al., 2007). Unlike single image based recognition, the test set in image set classification consists of a number of images belonging to the same class. The task is to assign a collective label to all the images in the test set. Similarly, the training or gallery data is also composed of one or more image sets for each class and each image set consists of different images of the same class.

Image set classification can provide solution to many of the shortcomings of single image based recognition methods. The extra information (poses and views) contained in an image set can be used to overcome a wide range of appearance variations such as non-rigid deformations, viewpoint changes, occlusions, different background and illumination variations. It can also help to take better decisions in

---

*Department of Computer Science and Software Engineering, The University of Western Australia. uzair.nadeem@research.uwa.edu.au

†Department of Computer Science and Software Engineering, The University of Western Australia. afaq.shah@uwa.edu.au

‡Department of Computer Science and Software Engineering, The University of Western Australia. mohammed.bennamoun@uwa.edu.au

§Department of Electrical, Electronics and Computer Engineering, The University of Western Australia. roberto.togneri@uwa.edu.au

¶School of Engineering and Information Technology, Murdoch University. f.sohel@murdoch.edu.au

the presence of noise, distortions and blur in images. These attributes render image set classification as a strong candidate for many real life applications e.g., surveillance, video-based face recognition, object recognition and person re-identification in a network of security cameras (Hayat et al., 2015).

The image set classification techniques proposed in the literature can broadly be classified into three categories. **(i)** The first category, known as parametric methods, uses statistical distributions to model an image set (Arandjelovic et al., 2005). These techniques calculate the closeness between the distributions of training and test image sets. These techniques however, fail to produce good results in the absence of a strong statistical relationship between the training and the test data. **(ii)** Contrary to parametric methods, non-parametric techniques do not model image sets with statistical distributions. Rather, they represent image sets as linear or nonlinear subspaces (Wang et al., 2012; Ortiz et al., 2013; Yang et al., 2013; Zhu et al., 2013). **(iii)** The third type of techniques use machine learning to classify image sets (Hayat et al., 2014, 2015; Lu et al., 2015).

Most current image set classification techniques have been evaluated on video-based face recognition. However, the results cannot be generalized to real life surveillance scenarios, because the images obtained by surveillance cameras are usually of low resolution and contrast, and contain blur and noise. Moreover, in surveillance cases, backgrounds in training and testing are usually different. The training set usually consists of high resolution mug shots in indoor studio environments, while the test set contains noisy images in highly uncontrolled environments. Another challenge is that not enough data is normally available for testing and training, e.g., even if there is a video of a person to be identified, captured by a surveillance camera, there may be very few frames in which the face of that person is recognizable. In addition to humans, it may also be required to recognize various objects in a surveillance scenario.

The current state-of-the-art techniques for image set classification are learning-based and they require a lot of training data to achieve good results. However, in practical scenarios (e.g., to identify a person captured in a surveillance video from a few mugshots)

very little training data is usually available per class. Moreover, the current techniques are not able to cope effectively with the challenges, e.g., of low resolution images. Image set classification techniques usually require hand-crafted features and parameter tuning to achieve good results. Computational time is another constraint as decisions for surveillance and security are required in real-time. Moreover, it is difficult to add new classes in methods that require training, as this will require the re-training on all of the data. An ideal image set classification technique should be able to distinguish image sets with less gallery data and should be capable of coping with the challenges of low resolution, noise and blur. The technique should also work in real time with minimal or no training (or parameter tuning) with the capability to easily add new classes.

This paper proposes a novel non-parametric approach for image set classification. The proposed technique is based on the concept of image projections on subspaces using Linear Regression Classification (LRC) (Naseem et al., 2010). LRC uses the principle that images of the same class lie on a linear subspace (Belhumeur et al., 1997; Basri and Jacobs, 2003). In our proposed technique, we used downsampled, raw grayscale images of gallery set of each category to form a subspace in a high dimensional space. Our technique does not involve any training. At test time, each test image is represented as a linear combination of images in each gallery image set. Regression model parameters are estimated for each image of the test image set using a least squares based solution. Due to the least squares solution, the proposed technique can be susceptible to the problem of singular matrix or singularity. This occurs when the matrix is rank deficient i.e., the rank of the matrix is less than the number of rows. We used perturbation to overcome this problem. The estimated regression model is used to calculate the projection of the test image on the gallery subspace. The residual obtained for the test image projection is used as the distance metric. Three different strategies: majority voting, nearest neighbour classification and weighted voting, are evaluated to decide for the class of the test image set. The block diagram of the proposed technique is shown in Figure 1. We

carried out a rigorous performance evaluation of the proposed technique for the tasks of real life surveillance scenarios on the SCFace Dataset (Grgic et al., 2011), FR_SURV Dataset (Rudrani and Das, 2011) and Choke Point Dataset(Wong et al., 2011). We also evaluated our algorithm for the task of video-based faced recognition on three popular image set classification datasets: CMU Motion of Body (CMU MoBo) Dataset (Gross and Shi, 2001), YoutubeCelebrity (YTC) Dataset (Kim et al., 2008) and UCSD/Honda Dataset (Lee et al., 2003). We used the Washington RGBD Dataset (Lai et al., 2011) and ETH-80 dataset (Leibe and Schiele, 2003) for experiments on object recognition. The experiments are carried out at different resolution and for different sizes of gallery and test sets to validate the performance of the proposed technique for low resolution and limited data scenarios. We also provide comparison with other prominent image set classification algorithms.

The main contribution of this paper is a novel image set classification technique that is able to produce a superior classification performance, especially under the constraints of low resolution and limited gallery and test data. Our technique does not require any training and can easily be generalized across different datasets. A preliminary version of this work for video-based face recognition appeared in Shah et al. (2017). However, the experiments in Shah et al. (2017) were mainly geared towards video-based face recognition datasets at normal resolutions and did not evaluate the ability of the proposed technique for real life surveillance scenarios e.g., situations in which the gallery and the test data have been acquired under significantly different conditions such as, camera noise and scale (i.e., distance of the imaging device from the object). This paper extends Shah et al. (2017) by improving the decision making process and provides three different strategies for the estimation of the class of the test image set. This paper also provides a more detailed explanation of the proposed technique and adds extensive experiments to demonstrate the superior capability of the proposed technique to perform effectively using a limited number of low resolution images. The experiments are performed at three different resolutions to test the performance of our technique under the changes

in resolution. To the best of our knowledge, this is the first implementation of image set classification for real life surveillance scenarios. We also evaluated the proposed technique for the tasks of set-based object classification. The effect of variations in the resolution and the size of the gallery, and the test data, on the accuracy and the computational efficiency of the proposed technique is also presented, along with the comparison with other techniques. The performance of the proposed technique on three different types of datasets (surveillance, video-based face recognition and set-based object recognition) also establishes the generalization capabilities of the technique. Also in Shah et al. (2017), the experimental results of the compared techniques were extracted from Hayat et al. (2015). However, due to variations in the testing protocols and in the data used for experiments, there are significant differences in the results. This issue is also mentioned by Chen (2014). Due to this, it has become a common practice in image set classification works to directly compare with the authors' implementation of the compared methods (e.g., in Yang et al. (2013); Zhu et al. (2013); Hayat et al. (2014, 2015); Lu et al. (2015)) to present a fair comparison. In this paper we have adopted the same approach and directly compared our technique against the original implementations as is provided by the respective authors.

The rest of this paper is organized as follows. Section 2 presents an overview of the related work. The proposed technique is discussed in Section 3. Section 4 reports the experimental conditions and a detailed evaluation of our results, compared to the state-of-the-art approaches. The comparison of the computational time of the proposed technique with other methods is presented in the Section 5. The paper is concluded in Section 6.

2   Related Work

Image set classification methods can be classified into three broad categories: parametric, non-parametric and machine learning based methods, depending on how they represent the image sets. The parametric methods represent image sets with a statistical distri-
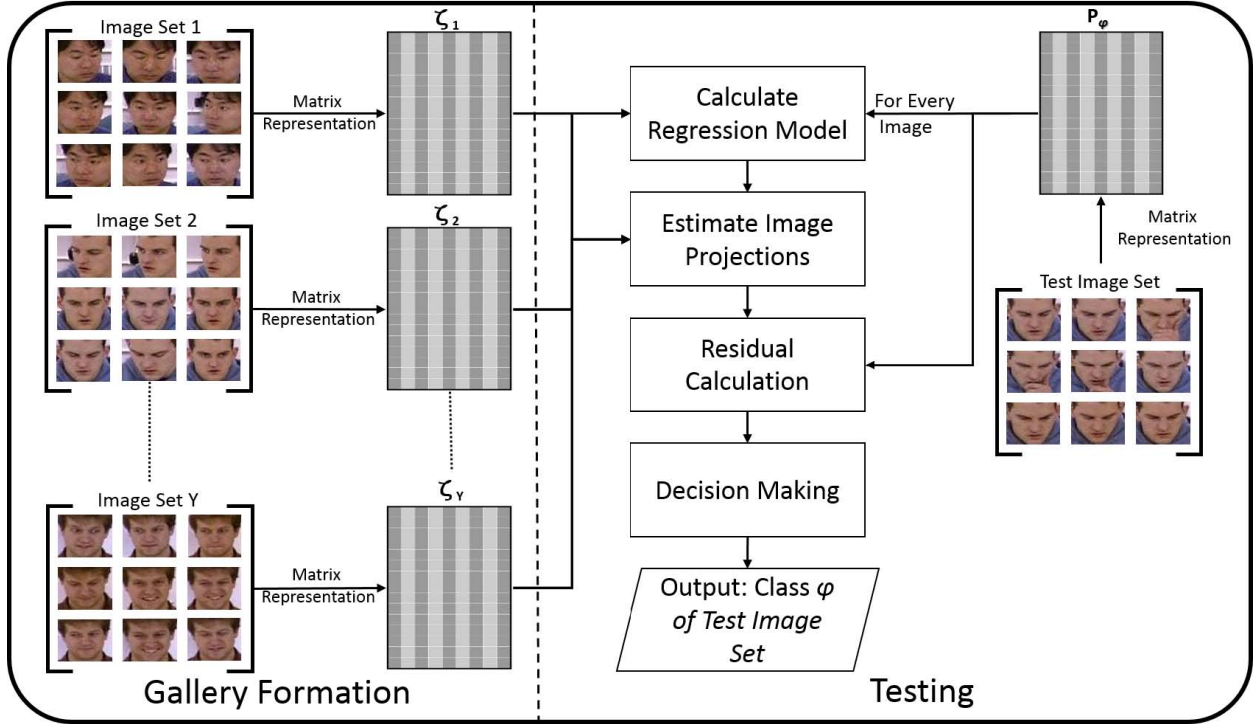
Figure 1: A block diagram of the proposed technique. Each gallery image set is converted to a matrix and is treated as a subspace. Each test image is projected on the gallery subspaces. The residuals, obtained by the difference of the original and the projected images, are used to decide on the class of the test image set.

bution model (Arandjelovic et al., 2005) and then use different approaches, such as KL-divergence, to calculate the similarity and differences between the estimated distribution models of gallery and test image sets. Such methods have the drawback that they assume a strong statistical relationship between the image sets of the same class and a weak statistical connection between image sets of different classes. However, this assumption does not hold many a times.

On the other hand, non-parametric methods model image sets by exemplar images or as geometric surfaces. Different works have used a wide variety of metrics to calculate the set to set similarity. Wang et al. (2008) use the Euclidean distance between the mean images of sets as the similarity metric. Cevikalp and Triggs (2010) adaptively learn samples

from the image sets and use affine hull or convex hull to model them. The distance between image sets is called Convex Hull Image Set Distance (CHISD) in the case of convex hull model, while in the case of affine hull model, it is termed as the Affine Hull Image Set Distance (AHISD). Hu et al. (2012) calculate the mean image of the image set and affine hull model to estimate the Sparse Approximated Nearest Points (SANP). The calculated Nearest Points determine the distance between the training and the test image sets. Chen et al. (2012) estimate different poses in the training set and then iteratively train separate sparse models for them. Chen et al. (2013) use a non-linear kernel to improve the performance of Chen et al. (2012). The non-availability of all the poses in all videos hampers the final clas-

4

sification accuracy. Some non-parametric methods e.g., (Wang et al., 2008; Wang and Chen, 2009; Harandi et al., 2011; Wang et al., 2012), represent the whole image set with only one point on a geometric surface. The image set can also be represented either by a combination of subspaces or on a complex non-linear manifold. In the case of subspace representation, the similarity metric usually depends on the smallest angles between any vector in one subspace and any other vector in the other subspace. Then the sum of the cosines of the smallest angles is used as the similarity measure between the two image sets. In the case of manifold representation of image sets, different metrics are used for distance, e.g., Geodesic distance (Turaga et al., 2011; Hayat and Bennamoun, 2014), the projection kernel metric on the Grassman manifold (Hamm and Lee, 2008), and the log-map distance metric on the Lie group of Riemannian manifold (Harandi et al., 2012). Different learning techniques based on discriminant analysis are commonly used to compare image sets on the manifold surface such as Discriminative Canonical Correlations (DCC) (Kim et al., 2007), Manifold Discriminant Analysis (MDA) (Wang and Chen, 2009), Graph Embedding Discriminant Analysis (GEDA) (Harandi et al., 2011) and Covariance Discriminative Learning (CDL) (Wang et al., 2012). Chen (Chen, 2014) assumes a virtual image in a high dimensional space and used the distance of the training and test image sets from the virtual image as a metric for classification. Feng et al. (2016) extends the work of Chen (2014) by using the the distance between the test set and unrelated training sets, in addition to the distance between the test image set and the related training set. However, these methods can only work for very small image sets due to the limitation that the dimension of the feature vectors should be much larger than the sum of the number of images in the gallery and the test sets.

Liu et al. (2018) use the relationship between image sets to model the inter-set and intra-set variations. They learn Group Collaborative Representations (GCRs) using a self-representation learning method, to model the probe sets using all the gallery data. Hayat et al. (2017) use a subset of gallery images and query set to train a binary SVM to learn a representation of the query set. This representation is then used to determine the similarity between the query and the gallery sets. However, their method is computationally expensive at the test time and it is not suitable for small query or gallery sets, since SVM is not able to learn a good representation in such a case. Lu et al. (2015) use deep learning to learn non-linear models and then apply discriminative and class specific information for classification. Hayat et al. (2014, 2015) propose an Adaptive Deep Network Template (ADNT) which is a deep learning based method. They use autoencoders to learn class specific non-linear models of the training sets. Instead of random initialization, Gaussian Restricted Boltzmann Machine (GRBM) is used to initialize the weights of the autoencoders. At test time, the class specific autoencoder models are used to reconstruct each image of the test set, and then the reconstruction error is used as a classification metric. ADNT has been shown to achieve good classification accuracies, but it requires the fine-tuning of several parameters and the extraction of hand crafted LBP features to achieve a good performance. Learning-based techniques are computationally expensive and require a large amount of training data to produce good results.

Our proposed technique estimates the linear regression model parameters for each image in the test set and then reconstructs test images from the gallery subspaces. Our technique can cope well with the challenges of low resolution, noise and small gallery and test data, and does not have any constraints on the number of images in the test set. Moreover, it does not involve any feature extraction or training and can produce state of the art results using only the raw images. At the same time our technique is computationally efficient and can classify image sets much faster than the learning based techniques both at the training and the test times (Section 5). It can also effectively utilize parallel processing or GPU computation to further reduce the computational time.
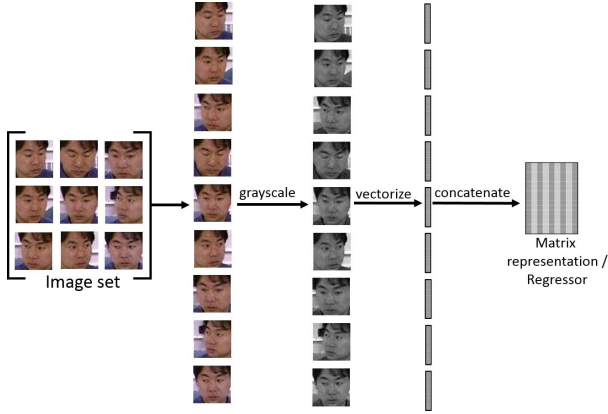
Figure 2: A representation of the regressor formation in the proposed technique.

3 Proposed Technique

In this paper, we adopt the following notation: bold capital letters denote **MATRICES**, bold lower letters show **vectors**, while normal letters are for *scalar numbers*. We use $[i]$ to denote the set of integer numbers from 1 to i, for a given integer i.

Consider the problem to classify the unknown class of a test image set as one of the $[Y]$ classes. Let the gallery image set of each unique class $[Y]$ contain N_ζ images. Each image in the gallery image sets is downsampled from the initial resolution of $a \times b$ to the resolution of $c \times d$ and are converted to grayscale to be represented as $\boldsymbol{G_y^i} \in \mathbb{R}^{c \times d}, \forall y \in [Y]$ and $\forall i \in [N_\zeta]$. We concatenate all the columns of each image to obtain a feature vector such that $\boldsymbol{G_y^i} \in \mathbb{R}^{c \times d} \to \boldsymbol{\varsigma_y^i} \in \mathbb{R}^{\tau \times 1}$, where $\tau = cd$. We do not use any feature extraction technique and use raw grayscale pixels as the features. Next, an image set matrix $\boldsymbol{\zeta_y}$ is formed for each class in $[Y]$ by horizontally concatenating the image vectors of class y.

$$\boldsymbol{\zeta_y} = [\boldsymbol{\varsigma_y^1 \varsigma_y^2 \varsigma_y^3 ... \varsigma_y^{N_\zeta}}] \in \mathbb{R}^{\tau \times N_\zeta}, \quad \forall y \in [Y] \qquad (1)$$

This follows the principle that patterns from the same class form a linear subspace (Basri and Jacobs, 2003). Each vector $\boldsymbol{\varsigma_y^i}, \forall i \in [N_\zeta]$, of the matrix $\boldsymbol{\zeta_y}$, spans a subspace of $\mathbb{R}^{\tau \times 1}$. Thus, each class in $[Y]$ is rep-

resented by a vector subspace $\boldsymbol{\zeta_y}$ called the *regressor* for class y. A representation of the regressor formation is shown in 2.

Let φ be the unknown class of the test image set with N_P number of images. In the same way as the gallery images, images of the test image set are downsampled to the resolution of $c \times d$ and converted to grayscale to be represented as $\boldsymbol{T_\varphi^j} \in \mathbb{R}^{c \times d}$ where φ is the unknown class of the test image set and $j \in [N_P]$. Each downsampled image is converted to a vector through column concatenation such that $\boldsymbol{T_\varphi^j} \in \mathbb{R}^{c \times d} \to \boldsymbol{\rho_\varphi^j} \in \mathbb{R}^{\tau \times 1}$, where $\tau = cd$. A matrix $\boldsymbol{P_\varphi}$ is formed for the test image set by the horizontal concatenation of image vectors $\boldsymbol{\rho_\varphi^j}, \forall j \in [N_P]$.

$$\boldsymbol{P_\varphi} = [\boldsymbol{\rho_\varphi^1 \rho_\varphi^2 \rho_\varphi^3 ... \rho_\varphi^{N_P}}] \in \mathbb{R}^{\tau \times N_P} \qquad (2)$$

If $\boldsymbol{P_\varphi}$ belongs to the yth class, then its vectors should span the same subspace as the image vectors of the gallery set $\boldsymbol{\zeta_y}$. In other words, image vectors of test set $\boldsymbol{P_\varphi}$ should be represented as a linear combination of the gallery image vectors from the same class. Let $\boldsymbol{\theta_y^j} \in \mathbb{R}^{N_\zeta \times 1}$ be a vector of parameters. Then the image vectors of test set can be represented as:

$$\boldsymbol{\rho_\varphi^j} = \boldsymbol{\zeta_y \theta_y^j}, \quad \forall j \in [N_P], \ \forall y \in [Y] \qquad (3)$$

The number of features must be greater than or equal to the number of images in each gallery set i.e., $\tau \geq N_\zeta$ in order to calculate a unique solution for Equation (3) by the Least Squares method (Friedman et al., 2001; Ryan, 2008; Seber and Lee, 2012). However, even if the condition of $\tau \geq N_\zeta$ holds, it is possible for one or more rows of the regressor $\boldsymbol{\zeta_y}$ to be linearly dependent on one another, which renders the regressor matrix $\boldsymbol{\zeta_y}$ to be singular. In this case, the regressor $\boldsymbol{\zeta_y}$ is called a rank deficient matrix since the rank of $\boldsymbol{\zeta_y}$ is less than the number of rows. In this scenario, it is not possible to use the least squares solution to estimate a unique parameter vector $\boldsymbol{\theta_y^j}$. The singularity of the regressor $\boldsymbol{\zeta_y}$ can be removed by perturbation (Wang et al., 2012). In the case of a singular matrix, a small perturbation matrix $\boldsymbol{\varepsilon}$ with uniform random values in the range $-0.5 \leq \epsilon \leq +0.5$ was added to the regressor $\boldsymbol{\zeta_y}$:

$$\widetilde{\boldsymbol{\zeta_y}} = \boldsymbol{\zeta_y} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \in \mathbb{R}^{\tau \times N_\zeta}, \quad \forall \epsilon \in \boldsymbol{\varepsilon} : \ |\epsilon| \leq 0.5 \quad (4)$$

Table 1: Average percentage classification accuracies for the task of surveillance on SCface (Grgic et al., 2011), FR_SURV (Rudrani and Das, 2011) and Choke Point (Wong et al., 2011) datasets.

Datasets→	SC_Face			FR_SURV			Choke Point		
Methods↓ \ Resolutions→	10×10	15×15	20×20	10×10	15×15	20×20	10×10	15×15	20×20
AHISD (Cevikalp and Triggs, 2010)	30.76	40.76	40	31.37	33.33	35.29	52.27	54.54	55.08
RNP(Yang et al., 2013)	29.23	36.92	36.15	27.45	33.33	35.29	46.84	48.82	49.27
SSDML (Zhu et al., 2013)	9.23	15.38	13.85	11.76	7.84	7.84	53.82	47.73	48.34
DLRC(Chen, 2014)	36.92	39.23	38.46	5.88	9.8	3.92	41.79	45.35	45.40
DRM (Hayat et al., 2015)	2.31	6.15	4.62	5.88	9.8	9.8	32.54	50.00	55.53
PLRC (Feng et al., 2016)	24.62	25.38	26.15	37.25	35.29	35.29	53.85	56.19	56.83
PDL (Wang et al., 2017)	8.46	9.23	9.23	23.53	27.45	27.45	38.32	41.31	42.70
DARG (Wang et al., 2018)	0.77	2.31	1.54	9.80	9.80	11.76	22.75	30.11	30.19
Ours (MV)	41.54	46.92	42.31	**45.10**	49.02	45.10	55.11	54.95	56.34
Ours (NN)	41.54	46.92	42.31	39.22	49.02	**47.06**	55.99	56.68	56.98
Ours (EWV)	**44.62**	**48.46**	**46.15**	41.18	**51**	**47.06**	**56.42**	**56.95**	**57.22**

The regressor $\boldsymbol{\zeta_y}$ is perturbed while its values are still in the range of 0 to 255. Therefore, the maximum possible change in the value of any pixel is 0.5. Now, the vector of parameters $\boldsymbol{\theta_y^j}$ can be estimated for the test image vector $\boldsymbol{\rho_\varphi^j}$ and regressor $\widetilde{\boldsymbol{\zeta_y}}$ by the following equation:

$$\boldsymbol{\theta_y^j} = (\widetilde{\boldsymbol{\zeta_y}}'\widetilde{\boldsymbol{\zeta_y}})^{-1}\widetilde{\boldsymbol{\zeta_y}}'\boldsymbol{\rho_\varphi^j}, \quad \forall j \in [N_P], \ \forall y \in [Y] \quad (5)$$

where $\widetilde{\boldsymbol{\zeta_y}}'$ is the transpose of $\widetilde{\boldsymbol{\zeta_y}}$. Next, the estimated vector of parameters $\boldsymbol{\theta_y^j}$ and the regressor $\widetilde{\boldsymbol{\zeta_y}}$ are used to calculate the projection of the image vector $\boldsymbol{\rho_\varphi^j}$ on the yth subspace. The projection $\widehat{\boldsymbol{\rho_y^j}}$ of the image vector $\boldsymbol{\rho_\varphi^j}$ on the yth subspace is calculated by the following equations:

$$\widehat{\boldsymbol{\rho_y^j}} = \widetilde{\boldsymbol{\zeta_y}}\boldsymbol{\theta_y^j}, \quad \forall j \in [N_P], \ \forall y \in [Y] \quad (6)$$

$$\widehat{\boldsymbol{\rho_y^j}} = \widetilde{\boldsymbol{\zeta_y}}(\widetilde{\boldsymbol{\zeta_y}}'\widetilde{\boldsymbol{\zeta_y}})^{-1}\widetilde{\boldsymbol{\zeta_y}}'\boldsymbol{\rho_\varphi^j} \quad (7)$$

The projection $\widehat{\boldsymbol{\rho_y^j}}$ can also be termed as the reconstructed image vector for $\boldsymbol{\rho_\varphi^j}$. Equations (6) and (7) are useful to estimate the image projections in online mode i.e., when all the test data is not available simultaneously. However, for an offline classification scenario, the matrix implementation of the above problem can be efficiently used to decrease the processing time. Let $\boldsymbol{\Theta_y} \in \mathbb{R}^{N_\zeta \times N_P}$ be a matrix of parameters. Then the test image set matrix $\boldsymbol{P_\varphi}$ can be represented as the combination of $\boldsymbol{\Theta_y}$ and the gallery set matrix $\widetilde{\boldsymbol{\zeta_y}}$:

$$\boldsymbol{P_\varphi} = \widetilde{\boldsymbol{\zeta_y}}\boldsymbol{\Theta_y}, \quad \forall y \in [Y] \quad (8)$$

$\boldsymbol{\Theta_y}$ can be calculated by using the least square estimation using the following equations:

$$\boldsymbol{\Theta_y} = (\widetilde{\boldsymbol{\zeta_y}}'\widetilde{\boldsymbol{\zeta_y}})^{-1}\widetilde{\boldsymbol{\zeta_y}}'\boldsymbol{P_\varphi}, \quad \forall y \in [Y] \quad (9)$$

The projections of the test image vectors on the yth subspace can be calculated as follows:

$$\widehat{\boldsymbol{P_y}} = \widetilde{\boldsymbol{\zeta_y}}\boldsymbol{\Theta_y}, \quad \forall y \in [Y] \quad (10)$$

$$\widehat{\boldsymbol{P_y}} = \widetilde{\boldsymbol{\zeta_y}}(\widetilde{\boldsymbol{\zeta_y}}'\widetilde{\boldsymbol{\zeta_y}})^{-1}\widetilde{\boldsymbol{\zeta_y}}'\boldsymbol{P_\varphi} \quad (11)$$

where $\widehat{\boldsymbol{P_y}} \in \mathbb{R}^{T \times N_P}$ is the matrix of projections of the image vectors for the test image set $\boldsymbol{P_\varphi}$ on the yth subspace. The difference between each test image $\boldsymbol{\rho_\varphi^j}$ and its projection $\widehat{\boldsymbol{\rho_y^j}}$, called residual, is calculated using the Euclidean distance:

$$r_y^j = \left\| \boldsymbol{\rho_\varphi^j} - \widehat{\boldsymbol{\rho_y^j}} \right\|_2, \quad \forall y \in [Y], \ \forall j \in [N_P] \quad (12)$$

Once the residuals are calculated we propose three strategies to reach the final decision for the unknown class φ of the test set:

3.1 Majority Voting (MV)

In majority voting, each image j of the test image set casts an equal weighted vote ϑ^j for the class y, whose regressor $\boldsymbol{\zeta_y}$ results in the minimum value of residual for that test image vector:

$$\vartheta^j = arg \min_y (r_y^j) \quad \forall j \in [N_p] \quad (13)$$

Then the votes are counted for each class and the class with the maximum number of votes is declared as the unknown class of the test image set.

$$\varphi_{MV} = mode([\vartheta^1 \vartheta^2 \vartheta^3 ... \vartheta^{N_P}]) \quad (14)$$

In the case of a tie between two or more classes, we calculate the mean of residuals for the competing classes, and the class with the minimum mean of residuals is declared as the output class of the test image set. Suppose classes $1, 2$ and 3 have the same number of votes then:

$$r_{y(mean)} = \underset{j}{mean}(r_y^j) \quad \forall y \in [1, 2, 3] \quad (15)$$

$$\varphi_{MV} = arg \min_y (r_{y(mean)}) \quad (16)$$

3.2 Nearest Neighbour Classification (NN)

In Nearest Neighbour Classification, we first find the minimum residual $r_{y(min)}$ across all classes in $[Y]$ for each of the test image vector.

$$r_{y(min)} = \min_j (r_y^j) \quad \forall y \in [Y] \quad (17)$$

Table 2: Average percentage classification accuracies for the task of video-based face recognition on UCSD/Honda (Honda)(Lee et al., 2003) dataset for different number of gallery and test set images.

No. of Images→	20-20			40-40			All-All		
Methods↓ \ Resolutions→	10×10	15×15	20×20	10×10	15×15	20×20	10×10	15×15	20×20
AHISD (Cevikalp and Triggs, 2010)	91.02	92.82	92.82	92.82	95.12	94.61	93.84	98.71	99.23
RNP(Yang et al., 2013)	94.35	95.12	95.12	95.64	97.43	96.41	95.64	99.48	99.48
SSDML (Zhu et al., 2013)	80.51	83.08	83.85	84.10	85.64	85.64	89.74	90.77	91.54
DLRC(Chen, 2014)	92.30	92.82	92.82	90.51	93.07	92.82	94.3	96.9	97.1
DRM (Hayat et al., 2015)	31.28	26.92	28.46	81.79	83.33	79.49	**100**	99.74	**100**
PLRC (Feng et al., 2016)	92.82	93.07	93.84	94.35	95.12	95.38	99.48	**100**	**100**
PDL (Wang et al., 2017)	88.97	90.77	91.28	96.41	96.41	94.87	98.72	99.74	99.49
DARG (Wang et al., 2018)	60.00	57.95	58.72	44.36	48.72	56.15	75.90	77.95	89.74
Ours (MV)	90.51	91.80	92.05	94.36	95.13	96.15	**100**	**100**	**100**
Ours (NN)	95.64	**96.15**	**95.90**	96.41	**97.95**	97.69	**100**	**100**	**100**
Ours (EWV)	**95.90**	**96.15**	**95.90**	**97.18**	**97.95**	**98.72**	**100**	**100**	**100**

Then, the class with the minimum value of $r_{y(min)}$ is declared as the output class of the test image set.

$$\varphi_{NN} = arg \min_{y}(r_{y(min)}) \qquad (18)$$

3.3 Exponential Weighted Voting (EWV)

In exponential weighted voting, each image $[N_P]$ in the test image set casts a vote ω_y^j to each class y in the gallery. We tested with different types of weights including the Euclidean distance, exponential of the Euclidean distance, the inverse of the Euclidean distance and the square of the inverse Euclidean distance. Empirically, it was found that the best performance was achieved when we used the exponential of the Euclidean distance as weights for the votes. Hence, the following equation defines the weight of

the vote ω_y^j of each image j :

$$\omega_y^j = e^{-\beta r_y^j}, \quad y \in [Y], \ \forall \ j \in [N_P] \qquad (19)$$

where β is a constant. The weights are accumulated for each gallery class $[Y]$ from each image of the test set:

$$\varpi_y = \sum_{j=1}^{N_P} \omega_y^j, \quad \forall \ y \in [Y] \qquad (20)$$

The final decision is ruled in the favour of the class y with the maximum accumulated weight from all the images ρ_φ^j of the test image set P_φ:

$$\varphi_{EWV} = arg \max_{y}(\varpi_y) \qquad (21)$$

3.4 Fast Linear Image Reconstruction

If the testing is done offline (i.e., all test data is available), then the processing time can substantially be reduced by using Equations (8), (9), (10) and (11)

instead of Equations (3), (5), (6) and (7). The computational efficiency of the proposed technique can further be improved by using Moore-Penrose pseudoinverse (Penrose, 1956; Stoer and Bulirsch, 2013) to estimate the inverse of the regressor matrix $\widetilde{\zeta_y}$ at the time of the gallery formation. This step significantly reduces the computations required at test time.

Table 7 shows that many fold gain in computational efficiency was achieved by the fast linear image reconstruction technique for the YouTube Celebrity dataset (Kim et al., 2008). Let $\widetilde{\zeta_y^{inv}}$ be the pseudoinverse of the regressor $\widetilde{\zeta_y}$ calculated at the time of gallery formation. Equation (8) can then be solved at the test time as:

$$\Theta_y = \widetilde{\zeta_y^{inv}} P_\varphi \qquad (22)$$

$$\widehat{P}_y = \widetilde{\zeta_y}(\widetilde{\zeta_y^{inv}} P_\varphi) \qquad (23)$$

4 Experiments and Analysis

We rigorously evaluated the proposed technique for the tasks of surveillance, video-based face recognition and object recognition. We evaluated our technique on three surveillance databases, namely SCFace (Grgic et al., 2011), FR_SURV (Rudrani and Das, 2011) and Choke Point (Wong et al., 2011) datasets. For video-based face recognition, we used the publicly available and challenging databases, namely Honda/UCSD Dataset (Lee et al., 2003), CMU Motion of Body Dataset (CMU MoBo) (Gross and Shi, 2001) and Youtube Celebrity Dataset (YTC) (Kim et al., 2008). Washington RGBD Object (Lai et al., 2011) and ETH-80 (Leibe and Schiele, 2003) datasets were used for the task of object recognition. We also performed experiments for limited number of images in the gallery and the test datasets to verify the performance of the proposed technique for the conditions where only scarce amount of data is available. All the experiments were performed at the resolutions of 20×20 pixels, 15×15 pixels and 10×10 pixels to establish the performance of our technique at extremely low resolution.

We compared our technique with several prominent image set classification methods. These techniques include the Affine Hull based Image Set Distance (AHISD) (Cevikalp and Triggs, 2010), Regularized Nearest Points (RNP) (Yang et al., 2013), Set to Set Distance Metric Learning (SSDML) (Zhu et al., 2013), Dual Linear Regression Classifier (DLRC) (Chen, 2014), Deep Reconstruction Models (DRM) (Hayat et al., 2015), Pairwise Linear Regression Classifier (PLRC) (Feng et al., 2016), Prototype Discriminative Learning (PDL) (Wang et al., 2017) and Discriminant Analysis on Riemannian Manifold of Gaussian Distributions (DARG) (Wang et al., 2018). We followed the standard experimental protocols. For the compared methods, we used the implementations provided by the respective authors. For all of the compared methods, the parameter settings which were recommended in the respective papers were used. For DARG (Wang et al., 2018), we kept 95% energy of the PCA features. Since DLRC (Chen, 2014) and PLRC (Feng et al., 2016) can work with only small datasets, we selected the number of training and testing images for these techniques depending on the best accuracy in each of the datasets.

4.1 Surveillance

4.1.1 SCFace Dataset

The SCface dataset (Grgic et al., 2011) contains static images of human faces. The main purpose of SCface dataset is to test face recognition algorithms in real world scenarios. This dataset presents the challenges of different training and testing environments, uncontrolled illumination, low resolution, less gallery and test data, head pose orientation and a large number of classes. Five commercially available video surveillance cameras of different qualities and various resolutions were used to capture the images of 130 individuals in uncontrolled indoor environments. Different quality cameras present a real world surveillance scenario for face recognition. Due to the large number of individuals, there is a very low probability of results obtained by pure chance. Images were taken under uncontrolled illumination conditions with outdoor light coming from a window as

the only source of illumination. Two of the surveillance cameras record both visible spectrum and IR night vision images, which are also included in the dataset. Images were captured from various distances for each individual. Another challenge in the dataset is the head pose i.e., the camera is above the subject's head to mimic a commercial surveillance systems while for the gallery images the camera is in front of the individual. For the gallery set, individuals were photographed with a high quality digital photographer's camera at close range in controlled conditions (standard indoor lighting, adequate use of flash to avoid shades, high resolution of images) to mimic the conditions used by law enforcement agencies. Different views were captured for each individual from -90 degrees to +90 degrees. Figure 3 clearly presents the difference in the quality of images in the gallery images and test images of four individuals in the SCFace dataset.

We used the images taken in controlled conditions for the gallery sets while all other images were used to form the test image sets. The faces in the images were detected using the Viola and Jones face detection algorithm (Viola and Jones, 2004). The gallery sets have a maximum of six images. The images were converted to grayscale and resized to 10×10, 15×15 and 20×20. Table 1 shows the classification accuracies of our technique compared to the other state of the art techniques. Our technique achieved the best classification accuracy on all resolutions. This can be attributed to the fact that there is not sufficient data available for training, so the techniques which depend heavily on data availability e.g., (Hayat et al., 2015) fail to produce good results.

4.1.2   FR_SURV Dataset

The FR_SURV dataset (Rudrani and Das, 2011) contains images of 51 individuals. The main purpose of this dataset is to test face recognition algorithms in long distance surveillance scenarios. This dataset presents the challenges of uncontrolled illumination, low resolution, blur, noise, less gallery and test data, head pose orientation and a large number of classes. Unlike most other datasets, the gallery and test images were captured in different conditions. The

gallery set contains twenty high resolution images of each individual captured from a close distance in controlled indoor conditions. The images contain slight head movements and the camera is placed parallel to the head position. There is a significant difference between the resolutions of the gallery and the test images. The test images were captured in outdoor conditions from a distance of 50m - 100m. The camera was placed at the height of around 20m - 25m. Due to the use of commercial surveillance camera and the large distance between the camera and the individual, the test images are of very low quality which makes it difficult to identify the individual. The difference in the gallery and test images of the FR_SURV dataset in Figure 4 shows the challenges of the dataset.

We used the indoor images for the gallery sets while the outdoor images were used to form the test image sets. Viola and Jones face detection algorithm (Viola and Jones, 2004) was used to detect faces in the images. The images were converted to grayscale and resized to 10×10, 15×15 and 20×20. The classification accuracies of our technique compared to the other image set classification techniques are shown in Table 1. Although the results of all techniques are low on this dataset, our technique achieved the best performance on all resolutions. The percentage gap in the performance of our technique and the second best technique is more than 15%. The techniques which require a large training data fail to produce any considerable results on this dataset.

4.1.3   Choke Point Dataset

The Choke Point dataset (Wong et al., 2011) consists of videos of 29 individuals captured using surveillance cameras. Three cameras at different position simultaneously record the entry or exit of a person from 3 viewpoints. The main purpose of this dataset was to identify or verify persons under real-world surveillance conditions. The cameras were placed above the doors to capture individuals walking through the door in a natural manner. This dataset presents the challenges of variations in illumination, background, pose and sharpness. The 48 video sequences of the dataset contain 64,204 face images at a frame rate

Figure 3: Faces detected by the Viola and Jones face detection algorithm of four individuals from the challenging SCFace Dataset (Grgic et al., 2011). The first three columns show the Gallery images while the remaining columns show the test images. The large difference between the quality of the gallery and test images is evident.

of 30 fps with an image resolution of 800×600 pixels. In all sequences, only one individual is present in the image at a time. The sequences contain different backgrounds depending on the door and whether the individual is entering or leaving the room. Overall, 25 individuals appear in 16 video sequences while 4 appear in 8 video sequences.

We used the Viola and Jones face detection algorithm (Viola and Jones, 2004) to detect faces in the images. The images were converted to grayscale and resized to 10×10, 15×15 and 20×20. As recommended by the authors of the dataset for the evaluation protocol (Wong et al., 2011), we used only images from the camera which captured the most frontal faces for each video sequence. This results in 16 image sets for 25 individuals and 8 image sets for 4 individuals. We randomly selected two image sets of each individual for the gallery set while the rest of the image sets were used as test image sets. The experiment was repeated ten times with a random selection of gallery and test image sets to improve the generalization of results. We achieved average classification accuracies of 71.23%, 73.85% and 72.75% for 10×10, 15×15 and 20×20 resolutions, respectively. In order to test the performance of image set classification methods under the challenges of limited gallery and

test data, we reduced the size of image sets to 20 by retaining the first 20 images of each image set. Table 1 shows the classification accuracies on the Choke Point Dataset. Our technique achieved the best classification accuracies compared to the other image set techniques.

4.2 Video-based Face Recognition

For video-based face recognition, we used the most popular image set classification datasets. For this section we also carried out extensive experiments using the full length image sets, in addition to the experiments which use limited gallery and test data, in order to demonstrate the suitability of our technique for both scenarios. In experiments using full length image sets, we reduced the gallery data, for our technique, whenever the number of images in the gallery set was larger than the number of pixels in the downsampled images. This ensures that the number of features is always larger than the number of images (see Section 3 for details) for our technique.

12

Figure 4: Random images of four individuals from the challenging FR_Surv Dataset (Rudrani and Das, 2011). The first five columns show the gallery images while the remaining columns show the test images. The difference in the quality of the gallery and the test images can clearly be observed.

4.2.1 UCSD/Honda Dataset

The UCSD/Honda Dataset (Lee et al., 2003) was developed to evaluate the face tracking and recognition algorithms. The dataset consists of 59 videos of 20 individuals. The resolution of each video sequence is 640×480. Of the 20 individuals, all but one have at least two videos in the dataset. All videos contain significant head rotations and pose variations. Moreover, some of the video sequences also contain partial occlusions in some frames. We followed the standard experimental protocol used in (Lee et al., 2003; Wang et al., 2008; Wang and Chen, 2009; Hu et al., 2012; Hayat et al., 2015). The face from each frame of videos was automatically detected using the Viola and Jones face detection algorithm (Viola and Jones, 2004). The detected faces were down sampled to the resolutions of 10×10, 15×15 and 20×20 and converted to grayscale. Histogram equalization was applied to increase the contrast of images. For the gallery image sets, we randomly selected one video of each individual. The remaining 39 videos were used as the test image sets. For our technique, we randomly selected a small number of frames to keep the number of gallery images considerably smaller than the number of pixels (refer to Section 3). To improve the consistency in scores we repeated the ex-

periment 10 times with a different random selection of gallery images, gallery image sets and test image sets. Our technique achieved 100% classification accuracy on all the resolutions while using a significantly less number of gallery images. We also carried out experiments for less gallery and test data by retaining the first 20 and first 40 images of each image set. Our technique achieved the best classification performance in all of the testing protocols. Table 2 summarizes the average identification rates of our technique compared to other image set classification techniques on the Honda dataset.

4.2.2 CMU MoBo Dataset

The original purpose of the CMU Motion of Body Dataset (CMU MoBo) was to advance biometric research on human gait analysis (Gross and Shi, 2001). It contains videos of 25 individuals walking on a treadmill, captured from six different viewpoints. Except the last one, all subjects in the dataset have four videos following different walking patterns. The four walking patterns are slow walk, fast walk, inclined walk and holding a ball while walking. Figure 5 shows random images of four individuals in the CMU MoBo dataset detected by the Viola and Jones face detection algorithm (Viola and Jones, 2004)

13

Table 3: Average percentage classification accuracies for the task of video-based face recognition on CMU MoBo (MoBo)(Gross and Shi, 2001) dataset for different sizes of gallery and test images.

No. of Images→	20-20			All-All		
Methods↓ \ Resolutions→	10×10	15×15	20×20	10×10	15×15	20×20
AHISD (Cevikalp and Triggs, 2010)	62.77	71.66	76.80	93.89	96.25	96.81
RNP(Yang et al., 2013)	66.25	72.5	77.63	92.36	95.69	97.36
SSDML (Zhu et al., 2013)	73.47	73.06	73.47	96.39	96.67	96.25
DLRC(Chen, 2014)	80.69	82.08	80.00	96.2	97.2	97.7
DRM (Hayat et al., 2015)	26.94	40.14	38.06	90	90.56	91.81
PLRC (Feng et al., 2016)	79.72	80.41	79.16	96.38	96.38	96.94
PDL (Wang et al., 2017)	70.28	72.50	74.17	98.33	98.33	98.47
DARG (Wang et al., 2018)	37.92	39.58	40.42	71.81	72.92	76.53
Ours (MV)	73.06	74.31	74.17	90.97	90.70	91.53
Ours (NN)	**80.97**	**83.06**	**82.92**	**98.61**	**99.31**	**99.03**
Ours (EWV)	80.56	**83.06**	**82.92**	**98.61**	**99.31**	**99.03**

Following the same procedure as previous works (Chen, 2014; Hayat et al., 2015; Feng et al., 2016), we used the videos of the 24 individuals which contain all the four walking patterns. Only the videos from the camera in front of the individual were used for image set classification. The frames of each video were considered as an image set. We followed the standard protocol (Wang et al., 2008; Cevikalp and Triggs, 2010; Hu et al., 2012; Hayat et al., 2015), and randomly selected the video of one walking pattern of each individual for the gallery image set. The remaining videos were considered as test image sets. As mentioned in Section 3, the number of images should be less than or equal to the number of pixels in the downsampled images. In practice, the number of images should be considerably less than the number of pixels. For our technique, we randomly selected a small number of frames from each gallery video to form the gallery image sets. The face from each frame was automatically detected using the Viola and Jones face detection algorithm (Viola and Jones, 2004). The images were resampled to the resolutions of 10×10, 15×15 and 20×20 and converted to grayscale. Histogram equalization was applied to increase the contrast of images. Unlike most other techniques, we used raw images for our technique and did not extract any features, such as LBP features used in

(Cevikalp and Triggs, 2010; Chen, 2014; Hayat et al., 2015; Feng et al., 2016). The experiments were repeated 10 times with different random selections of the gallery and the test sets. We also used different random selections of the gallery images in each round to make our testing environment more challenging. Our technique achieved the best performance on this dataset, with greater than 99% average classification accuracy. Then we reduced the number of images in each image set to the first 20. Our technique achieved the best classification accuracy in both experimental protocols. Table 3 provides the average accuracy of our technique along with a comparison with other methods.

4.2.3 Youtube Celebrity Dataset

The YoutubeCelebrity (YTC) Dataset (Kim et al., 2008) contains cropped clips from videos of 47 celebrities and politicians. In total, the dataset contains 1910 video clips. The videos are downloaded from the YouTube and are of very low quality. Due to their high compression rates, the video clips are very noisy and have a low resolution. This makes it a very challenging dataset. The Viola and Jones face detection algorithm (Viola and Jones, 2004) failed to detect any face in a large number of frames. Therefore,

Figure 5: Face images detected by the Viola and Jones face detection Algorithm of four individuals from CMU MoBo dataset (Gross and Shi, 2001).

the Incremental Learning Tracker (Ross et al., 2008) was used to track the faces in the video clips. To initialize the tracker, we used the initialization parameters[1] provided by the authors of (Kim et al., 2008). This approach, however, produced many tracking errors due to the low image resolution and noise. Although the cropped face region was not uniform across frames, we decided to use the automatically tracked faces without any refinement to make our experimental settings more challenging.

We followed the standard protocol which was also followed in (Wang et al., 2008; Wang and Chen, 2009; Wang et al., 2012; Hu et al., 2012; Hayat et al., 2015), for our experiments. The dataset was divided into five folds with a minimal overlap between the various folds. Each fold consists of 9 video clips per individual, to make a total of 423 video clips. Three videos of each individual were randomly selected to form the gallery set while the remaining six were used as six separate test image sets. All the tracked face images were resampled to the resolutions of 10×10, 15×15 and 20×20 and converted to grayscale. We applied histogram equalization to enhance the image contrast. Figure 6 shows histogram equalized images of four celebrities in the YouTube Celebrity dataset.

For the gallery formation in our technique, we randomly selected a small number of images from each of the three gallery videos per individual per fold in order to keep the number of gallery images less than the number of pixels (features). Our technique achieved the highest classification accuracy among all the methods, while using significantly less gallery data compared to the other methods. We also carried out experiments for less gallery and test data by selecting the first 20 images from each image set. If any video clip had less than 20 frames, all the frames of that video clip were used for gallery formation. In this way each gallery set had a maximum of 60 images while each test set had a maximum of 20 images. Our technique achieved the best classification accuracies for all resolutions.

Table 4 summarizes the average accuracies of the different techniques on the YouTube Celebrity dataset.

4.3 Object Recognition

4.3.1 Washington RGBD Object Dataset

The Washington RGB-D Object Dataset (Lai et al., 2011) is a large dataset, which consists of common household objects. The dataset contains 300 objects in 51 categories. The dataset was recorded at a res-

---

[1] http://seqam.rutgers.edu/site/index.php?option=com_content&view=article&id=64&Itemid=80

Table 4: Average percentage classification accuracies for the task of video-based face recognition on YouTube Celebrity (YTC) (Kim et al., 2008) dataset for different sizes of gallery and test image sets.

No. of images→	60-20			All-All		
Methods↓ \ Resolutions→	10×10	15×15	20×20	10×10	15×15	20×20
AHISD (Cevikalp and Triggs, 2010)	59.71	59.36	58.58	43.9	55.88	54.18
RNP(Yang et al., 2013)	58.22	57.65	56.59	65.46	65.74	65.46
SSDML (Zhu et al., 2013)	60.5	61.42	60.14	63.26	65.25	64.4
DLRC(Chen, 2014)	57.02	61.13	60.42	60.14	64.4	63.9
DRM (Hayat et al., 2015)	39.93	46.52	51.91	49.36	57.66	61.06
PLRC (Feng et al., 2016)	60.19	59.27	59.48	62.08	62.51	62.59
PDL (Wang et al., 2017)	53.05	52.77	53.40	54.61	54.47	54.82
DARG (Wang et al., 2018)	39.11	41.74	43.39	50.02	53.02	51.74
Ours (MV)	57.38	58.23	56.95	60.71	60.64	59.29
Ours (NN)	**60.78**	**61.49**	**61.35**	**66.45**	**66.81**	**66.46**
Ours (EWV)	59.65	60.71	60.78	65.25	66.24	65.89

Table 5: Average percentage classification accuracies for the task of object recognition on Washington RGB-D Object (Lai et al., 2011) and ETH-80 (Leibe and Schiele, 2003) datasets.

Datasets→	Washington RGBD Object			ETH-80		
Methods↓ \ Resolutions→	10×10	15×15	20×20	10×10	15×15	20×20
AHISD (Cevikalp and Triggs, 2010)	53.78	57.32	57.37	73.25	85.25	87.5
RNP(Yang et al., 2013)	52.72	55.60	57.17	74	74.25	79
SSDML (Zhu et al., 2013)	44.39	46.31	43.33	73.25	69	67.75
DLRC(Chen, 2014)	50.91	56.76	57.77	76.2	85.2	87.2
DRM (Hayat et al., 2015)	34.85	51.92	61.72	88.75	90.25	88.75
PLRC (Feng et al., 2016)	55.61	58.13	59.17	80.93	86.47	87.44
PDL (Wang et al., 2017)	43.23	48.03	46.92	76.50	82.75	83.50
DARG (Wang et al., 2018)	21.52	22.22	11.67	57.00	54.50	51.25
Ours (MV)	61.06	63.74	62.58	89.75	94.75	**96**
Ours (NN)	53.13	56.31	53.94	74.25	81	81
Ours (EWV)	**61.52**	**64.80**	**63.33**	**90.5**	**95.25**	95.75

Figure 6: Random histogram equalized, grayscale images of four celebrities from YouTube Celebrity dataset (Kim et al., 2008).

olution of 640×480 pixels. The objects were placed on a turntable and video sequences were captured at 30 fps for a complete rotation. A camera mounted at 3 different heights was used to capture 3 video sequences for each object so that the object is viewed from different angles.

We used the RGB images of the dataset in our experiments. Using the leave one out protocol defined by the authors of the dataset for object class recognition (Lai et al., 2011), we achieved average classification accuracies of 83.33%, 83.73% and 84.51% for 10×10, 15×15 and 20×20 resolutions respectively. To evaluate the proposed technique for less gallery and test data, 20 images were randomly selected from the three video sequences of each instance to form a single image set. Two image sets of each class were randomly selected to form the gallery set, while the remaining sets were used as test sets. The images were converted to grayscale and resized to 10×10, 15×15 and 20×20 resolutions. Table 5 shows the classification accuracies of the image set classification techniques on this dataset. Our technique achieved the best classification accuracies at all the resolutions.

4.3.2 ETH-80 Dataset

The ETH-80 dataset (Leibe and Schiele, 2003) contains eight object categories: apples, pears, tomatoes, cups, cars, horses, cows and dogs. There are ten different image sets in each object category. Each image set contains images of an object taken from 41 different view angles in 3D. We used the cropped images, which contain the object without any border area, provided by the dataset. Figure 7 shows few images of four of the classes from ETH-80 dataset.

The images were resized to the resolutions of 10×10, 15×15 and 20×20. We used grayscale raw images for our technique. Similar to (Kim et al., 2007; Wang and Chen, 2009; Wang et al., 2012; Hayat et al., 2015), five image sets of each object category were randomly selected to form the gallery set while the other five are considered to be independent test image sets. In order to improve the generalization of the results, the experiments were repeated 10 times for different random selections of gallery and test sets. Table 5 summarizes the results of our technique compared to other methods. We achieved superior classification accuracy compared to all the other techniques at all resolutions.
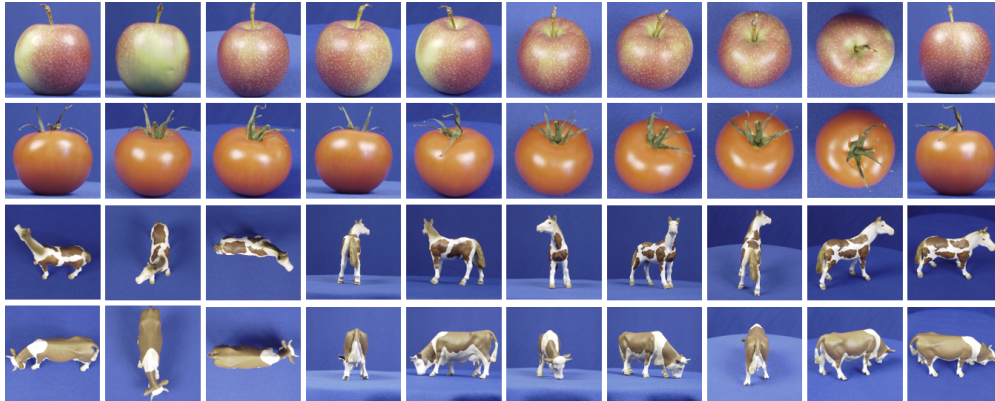
Figure 7: Random images of four of the classes from ETH-80 dataset. The resemblance between the images of apple (first row) and tomatoe (scond row), and between the images of horse (third row) and cow (fourth row) shows the challenge of inter-class similarity (Leibe and Schiele, 2003).

5 Computational Time Analysis

The fast approach of the proposed technique requires the least computational time, compared to other image set techniques. In contrast to the significant training times of SSDML (Zhu et al., 2013), DRM (Hayat et al., 2015) and DARG (Wang et al., 2018), the proposed technique does not require any training (Table 6). Hence, our technique is one of the fastest to set-up a face recognition system. Our technique is also the fastest at test time, especially for limited gallery and test data. The only method with a comparable efficiency at test time is PDL (Wang et al., 2017). Table 7 shows the testing times that are required per fold for the YouTube Celebrity dataset for various methods using a Core i7 3.40GHz CPU with 24 GB RAM. Although our technique reconstructs each image of the test image set from all the gallery image sets, due to its efficient matrix representation (refer to Section 3 and Section 3.4), we achieved a timing efficiency that is superior to the other methods. Moreover, due to the inherent nature of our technique it is highly parallelizable. It can be effectively deployed on embedded systems. The timing efficiency of the proposed technique can further be improved through parallel processing or GPU computation.

6 Conclusion and Future Direction

In this paper, we have proposed a novel image set classification technique, which exploits linear regression for efficient image set classification. The images from the test image set are projected on linear subspaces defined by down sampled gallery image sets. The class of the test image set is decided using the residuals from the projected test images. Extensive experimental evaluations for the tasks of surveillance, video-based face and object recognition, demonstrate the superior performance of our technique under the challenges of low resolution, noise and less gallery and test data, commonly encountered in surveillance settings. Unlike other techniques, the proposed technique has a consistent performance for the different tasks of surveillance, video-based face recognition and object recognition. The technique requires less gallery data compared to other image set classification methods and works effectively at very low resolution.

18

Table 6: Training time in seconds per fold required by different methods for the task of video-based face recognition on YouTube Celebrity (YTC) (Kim et al., 2008) dataset for different sizes of training and test image sets. NR shows that the method does not require training. * Includes time required to calculate LBP features.

No. of images→	60-20			All-All		
Methods↓ \ Resolutions→	10×10	15×15	20×20	10×10	15×15	20×20
AHISD (Cevikalp and Triggs, 2010)	NR	NR	NR	NR	NR	NR
RNP(Yang et al., 2013)	NR	NR	NR	NR	NR	NR
SSDML (Zhu et al., 2013)	9.74	18.89	21.61	101.69	193.37	278.63
DLRC(Chen, 2014)	NR	NR	NR	NR	NR	NR
DRM* (Hayat et al., 2015)	2194.28	2194.36	2194.40	4149.87	4157.23	4164.63
PLRC (Feng et al., 2016)	NR	NR	NR	NR	NR	NR
PDL (Wang et al., 2017)	5.10	5.18	5.23	60.33	60.71	62.18
DARG (Wang et al., 2018)	7.24	7.40	7.65	80.45	109.08	165.22
Ours	NR	NR	NR	NR	NR	NR
Ours (Fast)	NR	NR	NR	NR	NR	NR

Table 7: Testing time in seconds per fold required by different methods for the task of video-based face recognition on YouTube Celebrity (YTC) (Kim et al., 2008) dataset for different sizes of training and test image sets. Top-2 testing time for each setting are shown in bold. * Includes time required to calculate LBP features.

No. of images→	60-20			All-All		
Methods↓ \ Resolutions→	10×10	15×15	20×20	10×10	15×15	20×20
AHISD (Cevikalp and Triggs, 2010)	22.78	30.09	33.01	63.39	233.37	568.34
RNP(Yang et al., 2013)	8.16	14.20	16.07	21.35	31.83	43.65
SSDML (Zhu et al., 2013)	17.26	31.75	32.06	183.36	371.79	498.54
DLRC(Chen, 2014)	6.6262	12.40	23.57	303.55	347.55	431.23
DRM* (Hayat et al., 2015)	89.15	89.46	89.69	381.82	386.65	391.97
PLRC (Feng et al., 2016)	27.91	61.92	104.62	203.44	248.00	299.58
PDL (Wang et al., 2017)	**2.48**	**2.84**	**2.90**	**5.61**	**6.46**	**6.51**
DARG (Wang et al., 2018)	9.30	11.67	12.25	110.87	255.11	478.33
Ours	4.75	9.71	14.27	6.55	13.60	25.37
Ours (Fast)	**0.52**	**1.28**	**1.89**	**2.22**	**4.61**	**9.06**

Our proposed technique can easily be scaled from processing one frame at a time (for live video acquisition) to the processing of all the test data at once (for a faster performance). It can also be quickly and easily deployed as it is very easy to add new classes in our technique (by just arranging images in a matrix) and it does not require any training or feature extraction. Also in the case of resource rich scenarios, our technique can easily be parallelized to further increase the timing efficiency. These advantages make our technique ideal for practical applications such as surveillance. For future research, we plan to integrate non-linear regression and automatic image selection to make our approach more robust to outliers and to further improve its classification accuracy.

acknowledgements

References

Arandjelovic O, Shakhnarovich G, Fisher J, Cipolla R, Darrell T (2005) Face recognition with image sets using manifold density divergence. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, vol 1, pp 581–588

Basri R, Jacobs DW (2003) Lambertian reflectance and linear subspaces. IEEE transactions on Pattern Analysis and Machine Intelligence 25(2):218–233

Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7):711–720

Cevikalp H, Triggs B (2010) Face recognition based on image sets. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2567–2573

Chen L (2014) Dual linear regression based classification for face cluster recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2673–2680

Chen YC, Patel VM, Phillips PJ, Chellappa R (2012) Dictionary-based face recognition from video. In: European Conference on Computer Vision, Springer, pp 766–779

Chen YC, Patel VM, Shekhar S, Chellappa R, Phillips PJ (2013) Video-based face recognition via joint sparse representation. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, IEEE, pp 1–8

Feng Q, Zhou Y, Lan R (2016) Pairwise linear regression classification for image set retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 4865–4872

Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning, vol 1. Springer series in statistics Springer, Berlin

Grgic M, Delac K, Grgic S (2011) Scface–surveillance cameras face database. Multimedia tools and applications 51(3):863–879

Gross R, Shi J (2001) The cmu motion of body (mobo) database

Hamm J, Lee DD (2008) Grassmann discriminant analysis: a unifying view on subspace-based learning. In: Proceedings of the 25th international conference on Machine learning, ACM, pp 376–383

Harandi MT, Sanderson C, Shirazi S, Lovell BC (2011) Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2705–2712

Harandi MT, Sanderson C, Wiliem A, Lovell BC (2012) Kernel analysis over riemannian manifolds for visual recognition of actions, pedestrians and textures. In: Applications of Computer Vision (WACV), 2012 IEEE Workshop on, IEEE, pp 433–439

Hayat M, Bennamoun M (2014) An automatic framework for textured 3d video-based facial expression recognition. IEEE Transactions on Affective Computing 5(3):301–313

Hayat M, Bennamoun M, An S (2014) Learning nonlinear reconstruction models for image set classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1907–1914

Hayat M, Bennamoun M, An S (2015) Deep reconstruction models for image set classification. IEEE transactions on Pattern Analysis and Machine Intelligence 37(4):713–727

Hayat M, Khan SH, Bennamoun M (2017) Empowering simple binary classifiers for image set based face recognition. International Journal of Computer Vision 123(3):479–498

Hu Y, Mian AS, Owens R (2012) Face recognition using sparse approximated nearest points between image sets. IEEE transactions on Pattern Analysis and Machine Intelligence 34(10):1992–2004

Kim M, Kumar S, Pavlovic V, Rowley H (2008) Face tracking and recognition with visual constraints in real-world videos. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1–8

Kim TK, Kittler J, Cipolla R (2007) Discriminative learning and recognition of image set classes using canonical correlations. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6):1005–1018

Lai K, Bo L, Ren X, Fox D (2011) A large-scale hierarchical multi-view rgb-d object dataset. In: Robotics and Automation (ICRA), 2011 IEEE International Conference on, IEEE, pp 1817–1824

Lee KC, Ho J, Yang MH, Kriegman D (2003) Video-based face recognition using probabilistic appearance manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, vol 1, pp I–313

Leibe B, Schiele B (2003) Analyzing appearance and contour based methods for object categorization. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, vol 2, pp II–409

Liu B, Jing L, Li J, Yu J, Gittens A, Mahoney MW (2018) Group collaborative representation for image set classification. International Journal of Computer Vision pp 1–26

Lu J, Wang G, Deng W, Moulin P, Zhou J (2015) Multi-manifold deep metric learning for image set classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1137–1145

Naseem I, Togneri R, Bennamoun M (2010) Linear regression for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(11):2106–2112

Ortiz EG, Wright A, Shah M (2013) Face recognition in movie trailers via mean sequence sparse representation-based classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 3531–3538

Penrose R (1956) On best approximate solutions of linear matrix equations. In: Mathematical Proceedings of the Cambridge Philosophical Society, Cambridge Univ Press, vol 52, pp 17–19

Ross DA, Lim J, Lin RS, Yang MH (2008) Incremental learning for robust visual tracking. International Journal of Computer Vision 77(1-3):125–141

Rudrani S, Das S (2011) Face recognition on low quality surveillance images, by compensating degradation. International Conference on Image Analysis and Recognition (ICIAR 2011, Part II); Burnaby, BC, Canada, Lecture Notes in Computer Science (LNCS) 6754/2011:212–221

Ryan TP (2008) Modern regression methods, vol 655. John Wiley & Sons

Seber GA, Lee AJ (2012) Linear regression analysis, vol 936. John Wiley & Sons

Shah SAA, Nadeem U, Bennamoun M, Sohel F, Togneri R (2017) Efficient image set classification using linear regression based image classification. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops

Stoer J, Bulirsch R (2013) Introduction to numerical analysis, vol 12. Springer Science & Business Media

Turaga P, Veeraraghavan A, Srivastava A, Chellappa R (2011) Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(11):2273–2286

Viola P, Jones MJ (2004) Robust real-time face detection. International Journal of Computer Vision 57(2):137–154

Wang R, Chen X (2009) Manifold discriminant analysis. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 429–436

Wang R, Shan S, Chen X, Gao W (2008) Manifold-manifold distance with application to face recognition based on image set. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1–8

Wang R, Guo H, Davis LS, Dai Q (2012) Covariance discriminative learning: A natural and efficient approach to image set classification. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 2496–2503

Wang W, Wang R, Shan S, Chen X (2017) Prototype discriminative learning for face image set classification. IEEE Signal Processing Letters 24(9):1318–1322

Wang W, Wang R, Huang Z, Shan S, Chen X (2018) Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets. IEEE Transactions on Image Processing 27(1):151–163

Wong Y, Chen S, Mau S, Sanderson C, Lovell BC (2011) Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp 74–81

Yang M, Zhu P, Van Gool L, Zhang L (2013) Face recognition based on regularized nearest points between image sets. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, IEEE, pp 1–7

Zhu P, Zhang L, Zuo W, Zhang D (2013) From point to set: Extend the learning of distance metrics. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2664–2671