Efficient Image Set Classification using Linear Regression based Image Reconstruction

S.A.A. Shah*†, U. Nadeem*†, M. Bennamoun†, F. Sohel‡, and R. Togneri†

†The University of Western Australia
‡Murdoch University

†{afaq.shah@, uzair.nadeem@research., mohammed.bennamoun@, roberto.togneri@} uwa.edu.au
‡f.sohel@murdoch.edu.au

Abstract

We propose a novel image set classification technique using linear regression models. Downsampled gallery image sets are interpreted as subspaces of a high dimensional space to avoid the computationally expensive training step. We estimate regression models for each test image using the class specific gallery subspaces. Images of the test set are then reconstructed using the regression models. Based on the minimum reconstruction error between the reconstructed and the original images, a weighted voting strategy is used to classify the test set. We performed extensive evaluation on the benchmark UCSD/Honda, CMU Mobo and YouTube Celebrity datasets for face classification, and ETH-80 dataset for object classification. The results demonstrate that by using only a small amount of training data, our technique achieved competitive classification accuracy and superior computational speed compared with the state-of-the-art methods.

1. Introduction

Image set classification is defined as the problem of recognition from multiple images [16]. In image set classification, the gallery or training set consists of one or more image sets for each class and each image-set contains multiple images of the same class [16]. The test set also contains a number of images of the same subject which are then matched with the training image sets by computing some similarity measure to find the identity of the test subject.

Compared with traditional single image based recognition, image set classification offers several advantages. For instance, image sets can effectively handle a wide variety of appearance variations within images including: viewpoint changes, occlusions, non-rigid deformation, variations in illumination and different backgrounds. Because of these characteristics, image set classification has been applied in many applications in biometrics including surveillance, video based face recognition and person re-identification in a network of security cameras [13]. Several image set classification techniques have been proposed in the literature. A few of these techniques, known as parametric methods [1], model image sets with certain statistical distributions and then calculate the simlarity between those distributions. However, these methods require a strong statistical relationship between the test and the training image sets to achieve good performance. As opposed to these methods, non-parametric methods represent image sets as linear or non-linear subspaces [21], [30], [33], [35]. These methods have shown promising results and are being actively researched.

In this paper, we propose a novel non-parametric approach for image set classification. The proposed technique is based on the concept of image reconstruction using Lin-

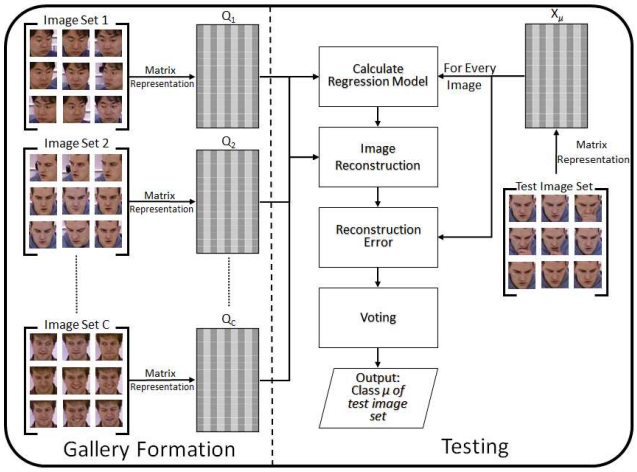

Figure 1. A block diagram of the proposed technique.

*The first two authors contributed equally to this work

ear Regression Classification (LRC) [20] and nearest subspace classification. LRC uses the concept that samples of an image category lie on a linear subspace [2], [3]. In our proposed technique, the gallery image set of each category forms a subspace in a high dimensional space while using the downsampled images of each gallery image set. At test time, each image in the test image set is represented as a linear combination of images in each gallery image set. A least squares based solution is used to estimate the regression model parameters for each image of the test image set. The estimated regression model is used to reconstruct the test image from the gallery subspace. The Euclidean distance between the actual test image and the reconstructed image is then used as the distance metric. Next, weighted voting is used where each image of the test image set casts a vote for each class in the gallery. Finally, the decision rules in favor of the class with the highest accumulated weight. Figure 1 shows the block diagram of the proposed technique. The performance of the proposed technique has been tested on four popular image set classification datasets, CMU Motion of Body (CMU MoBo) Dataset [10], Youtube Celebrity (YTC) Dataset [15] and UCSD/Honda Dataset [17] for face recognition, and ETH-80 dataset [18] for object recognition. We provide comparison with seventeen image set classification algorithms. The main contributions of this paper can be summarized as follows:

- A novel extension of LRC for image set classification, which is capable of producing state of the art results under the challenges of low resolution and less training data. The technique does not require any training and can easily be generalized across different datasets.
- Since LRC uses least squares solution, any technique using LRC is prone to the problem of singular matrix or singularity. This occurs when the rank is less than the number of rows in the regressor, a condition known as rank deficient matrix. While this problem is mostly ignored in the previous works of image set classification, we present practical and efficient solutions to overcome the problem of a rank deficient matrix. The solution is not limited to our technique and can be generalized to any method using LRC and least squares solutions.
- The techniques performing operations on each image of the test set are usually very slow and unsuitable for real time applications. On the other hand our technique uses an efficient matrix implementation of LRC to achieve the fastest test time compared to other image set classification methods.

The rest of this paper is organized as follows. An overview of related work is presented in Section 2. Section 3 discusses the proposed technique. Experimental results and detailed evaluation of the proposed technique against state-of-the-art approaches are presented in Section 4. The comparison of computational time of the proposed technique with other methods is presented in Section 5. The technique is compared with other latest image set classification methods in Section 6 and concluded in Section 7.

2. Related Work

Image set classification techniques can be categorized as parametric, non-parametric and deep learning based methods. The parametric methods [1] use a statistical distribution model to approximate an image set and then uses KL-divergence to measure the similarity between the two distribution models. Such methods, however, fail to produce good results in the case of a weak statistical relationship between the training and the test image sets.

For non-parametric methods, several different metrics are used to determine the set to set similarity. Wang et al. [31] use the Euclidean distance between the sets' mean as the similarity metric. Cevikalp and Triggs [4] present two models to learn set samples. The set to set distance using an affine hull model is called Affine Hull Image Set Distance (AHISD) while that using convex hull model is termed as the Convex Hull Image Set Distance (CHISD). Hu et al. [14] used the mean image of image set and affine hull model to calculate the Sparse Approximated Nearest Points (SANP) for image sets in order to determine the distance between the training image set and test image set. Chen et al. [6] iteratively trained separate models for different poses in the training set while enforcing sparsity constraints. Chen et al. [7] is an extension of [6], which uses a non-linear kernel to improve the performance. The non-availability of all the poses in all videos decreases final classification accuracy. Some non-parametric methods (e.g., [11], [29], [30], [31]) use a point on a geometric surface to represent the complete image set. The image set can also be represented either by a combination of linear subspaces or on a complex non-linear manifold. For linear subspaces, the cosine of the smallest angle between any vector in one subspace and any other vector in the other subspace is commonly used as the similarity metric between image sets.

Discriminant analysis is commonly used to represent image sets on the manifold surface e.g., Discriminative Canonical Correlations (DCC) [16], Manifold Discriminant Analysis (MDA) [29], Graph Embedding Discriminant Analysis (GEDA) [11] and Covariance Discriminative Learning (CDL) [30]. Chen [5] assumed a virtual image in a high dimensional space and used its distance from training and test image sets for classification. Feng et al. [8] extended the work of [5] by using the information which maximizes the distance between the test set and unrelated training sets. However, for these methods, the dimension of the feature vectors should be much larger than the combined number of images in the gallery and the test sets. Due to this lim-

itation, these methods only work for very small test sets. Zhang et al. [34] used constraint based learning and hashing function to model image sets in terms of binary codes. Lu et al. [19] used deep learning to learn non-linear models and exploit discriminative and class specific information for classification. Hayat et al. [12], [13] proposed a deep learning based approach called the Adaptive Deep Network Template (ADNT). In their technique, a deep autoencoder is used to define class-specific models for training sets. The weights of an autoencoder are initialized with a Gaussian Restricted Boltzmann Machine (GRBM). For classification, each image of the test set is reconstructed using a learnt class-specific model and the reconstruction error is used as a measure to identify the test image set. ADNT has been demonstrated to achieve state-of-the-art performance, but it relies on hand crafted LBP features and requires fine tuning of several parameters for achieving good performance. Moreover, deep learning methods require a large number of training images and are computationally expensive.

Our technique reconstructs images in the test image set using LRC from the gallery image matrix and is much faster than ADNT, both at training and test times. The proposed technique does not have any constraints on the number of images in the test set. Moreover, our technique can produce state of the art results using lower resolution raw images and much fewer training data, compared to other techniques.

3. Proposed Technique

Let N be the number of gallery images in each unique class C of the gallery set K_c. Each image is converted to grayscale and downsampled to a resolution of $a \times b$ to be represented as $k_c^n \in \mathbb{R}^{a \times b}$, where $c = 1, 2, 3, ..., C$ and $n = 1, 2, 3, ..., N$. Each gallery image is transformed through column concatenation to a vector such that $k_c^n \in \mathbb{R}^{a \times b} \rightarrow q_c^n \in \mathbb{R}^{T \times 1}$, where $T = ab$. Based on the concept that a linear subspace is formed by patterns from the same class [2], a class specific model Q_c is constructed for each class c by horizontally concatenating the image vectors of class c.

$$Q_c = [q_c^1 q_c^2 q_c^3 ... q_c^N] \in \mathbb{R}^{T \times N}, \quad c = 1, 2, 3, ..., C \tag{1}$$

In this way, each class c is represented by a vector subspace Q_c called the *regressor* for class c. Each vector $q_c^n,\ n = 1, 2, 3, ..., N,$ of the regressor Q_c spans a subspace of $\mathbb{R}^{T \times 1}$.

Let the problem be to classify the unknown class μ of a test image set Y_μ with M number of images, to one of the classes $c = 1, 2, 3, ..., C$. Similar to the gallery images, each image of the test image set is converted to grayscale and downsampled to the resolution of $a \times b$ to be represented as $y_\mu^m \in \mathbb{R}^{a \times b}$ where μ is the unknown class and $m = 1, 2, 3, ..., M$. Each downsampled image is transformed through column concatenation to a vector such that $y_\mu^m \in \mathbb{R}^{a \times b} \rightarrow x_\mu^m \in \mathbb{R}^{T \times 1}$, where $T = ab$. The image vectors $x_\mu^m, m = 1, 2, 3, ..., M$ are concatenated horizontally to create the test matrix X_μ

$$X_\mu = [x_\mu^1 x_\mu^2 x_\mu^3 ... x_\mu^M] \in \mathbb{R}^{T \times M}, \tag{2}$$

where μ is the unknown class. If X_μ belongs to the c^{th} class then it should be possible to represent the image vectors of X_μ as a linear combination of the gallery images from the same class i.e.,

$$x_\mu^m = Q_c \gamma_c^m,\ m = 1, 2, ..., M,\ c = 1, 2, ..., C \tag{3}$$

where $\gamma_c^m \in \mathbb{R}^{N \times 1}$ is a vector of parameters. For the unique solution of Equation (3) to exist, the condition $T \geq N$ must hold. Given that the condition holds, γ_c^m can be estimated for test image vector x_μ^m and regressor Q_c by using the least squares method [9], [25], [26]:

$$\gamma_c^m = (Q_c' Q_c)^{-1} Q_c' x_\mu^m,\ m = 1, 2, ..., M\ c = 1, 2, ..., C \tag{4}$$

where Q_c' is the transpose of Q_c. The image vector x_μ^m can be reconstructed for the class c using the parameters vector γ_c^m and the regressor Q_c:

$$\widehat{x}_c^m = Q_c \gamma_c^m,\ m = 1, 2, ..., M\ c = 1, 2, ..., C \tag{5}$$

$$\widehat{x}_c^m = Q_c (Q_c' Q_c)^{-1} Q_c' x_\mu^m \tag{6}$$

where $\widehat{x}_c^m$ is the reconstructed image vector for x_μ^m from the regressor Q_c. $\widehat{x}_c^m$ can also be interpreted as the projection of x_μ^m on the c^{th} subspace.

Instead of solving Equation (6) individually for each image vector x_μ^m, it can be formulated in the matrix form to efficiently utilize the computational power of modern computers:

$$X_\mu = Q_c \Gamma_c,\ c = 1, 2, ..., C \tag{7}$$

where $\Gamma_c \in \mathbb{R}^{N \times M}$ is a matrix of parameters. Γ_c can be calculated by using the least square estimation.

$$\Gamma_c = (Q_c' Q_c)^{-1} Q_c' X_\mu,\ c = 1, 2, ..., C \tag{8}$$

$$\widehat{X}_c = Q_c \Gamma_c,\ c = 1, 2, ..., C \tag{9}$$

$$\widehat{X}_c = Q_c (Q_c' Q_c)^{-1} Q_c' X_\mu \tag{10}$$

where $\widehat{X}_c \in \mathbb{R}^{T \times M}$ is the matrix of reconstructed image vectors for X_μ from the regressor Q_c. The reconstruction error between each test image x_μ^m and the reconstructed image $\widehat{x}_c^m$ is calculated using the Euclidean distance:

$$d_c^m = \left\| x_\mu^m - \widehat{x}_c^m \right\|_2,\ c = 1, 2, ..., C,\ m = 1, 2, ..., M \tag{11}$$

We tested different voting strategies. We empirically found that weighted voting consistently provides better performance on all the datasets. In weighted voting, each image m of the test image set casts a vote θ_c^m for each class

c to determine the class of the test image set X_μ. We experimented using the Euclidean distance, the inverse of the Euclidean distance, the square of inverse Euclidean distance and exponential of the Euclidean distance as weights. However, the best performance was achieved when using the exponential of the Euclidean distance in weighted voting. Hence, the weight of vote θ_c^m of each image m is defined by the following equation:

$$\theta_c^m = e^{-\alpha d_c^m},\ c = 1, 2, ..., C,\ m = 1, 2, ..., M \quad (12)$$

where α is a constant. The accumulated weight for each class c from each image of test set is given by:

$$\Theta_c = \sum_{m=1}^{M} \theta_c^m,\ c = 1, 2, ..., C \quad (13)$$

The class c which gets the maximum accumulated weight from all the images x_μ^m of the test image set X_μ is decided as the class of the test image set:

$$\mu = arg\ \max_c(\Theta_c)\ c = 1, 2, ..., C \quad (14)$$

Algorithm 1 provides the proposed image set classification technique.

Algorithm 1: The Proposed Image Set Classification Technique

Input : Gallery image sets K_c, where $c = 1, 2, 3, ...C$. Test image set Y_μ.
Output: Class μ of test image set Y_μ.
Gallery Formation:
for c *in* 1 *to* C **do**
  for n *in* 1 *to* N **do**
    $q_c^n \in \mathbb{R}^{T\times 1} = downsample\ images\ to\ a \times b$ and $vectorize$, where $T = ab$
  end
  $Q_c \in \mathbb{R}^{T\times N} = [q_c^1 q_c^2 q_c^3 ... q_c^N]$
end
Testing:
for m *in* 1 *to* M **do**
  $x_\mu^m \in \mathbb{R}^{T\times 1} = downsample\ images\ to\ a \times b$ and $vectorize$, where $T = ab$
end
$X_\mu \in \mathbb{R}^{T\times M} = [x_\mu^1 x_\mu^2 x_\mu^3 ... x_\mu^M]$
for c *in* 1 *to* C **do**
  for m *in* 1 *to* M **do**
    $\gamma_c^m = (Q_c' Q_c)^{-1} Q_c' x_\mu^m$
    $\widehat{x}_c^m = Q_c \gamma_c^m$
    $d_c^m = \sqrt{\sum^T ((x_\mu^m - \widehat{x}_c^m)^2)}$
    $\theta_c^m = e^{-\alpha d_c^m}$
  end
  $\Theta_c = \sum_{m=1}^{M} \theta_c^m$
end
$\mu = arg\ \max_c(\Theta_c)$

3.1. The Problem of Singularity

As mentioned before, for Equation (3) and Equation (7) to be well conditioned, the total number of pixels $T = ab$ in downsampled gallery image vectors q_c^n must be greater than or equal to the number of gallery images N in each regressor Q_c i.e., $T \geq N$. However, even if this condition holds, it is possible for regressor Q_c to be singular as one or more of the rows of Q_c may come out to be linearly dependent on other rows. In this case, regressor Q_c is called rank deficient due to the fact that $r < T$, where r is the rank of Q_c. Therefore, it is not possible to use Equation (4) and Equation (8) to calculate the parameters vector γ_c or parameters matrix Γ_c. In this paper we present two solutions for this problem:

3.1.1 Perturbation

The singularity of the regressor Q_c can be overcome by regularizing the regressor Q_c by adding a small perturbation term [30]. We empirically found that by adding a matrix ε with uniform random values in the range $-0.5 \leq \epsilon \leq +0.5$ removes the singularity of the regressor Q_c i.e.,

$$Q_c^* = Q_c + \varepsilon,\ \varepsilon \in \mathbb{R}^{T\times N} and\ \forall \epsilon \in \varepsilon,\ -0.5 \leq \epsilon \leq +0.5 \quad (15)$$

Note that Equation (15) is implemented before any preprocessing and the values in matrix Q_c are in the range of 0 to 255. In this way, the maximum possible change in the value of any pixel is 0.5. We observed that there was no deterioration in the classification accuracy when using this method.

3.1.2 Basic Solution using QR decomposition

In our second solution, we overcome the problem of singularity by computing a basic solution for Equation (3) or Equation (7) using QR decomposition [23], [27] of the regressor Q_c with the condition that the number of non-zero components in the solution vector $\gamma_c \leq r$, where r is the rank of the regressor Q_c. This method does not remove the singularity of the regressor Q_c, however, the results obtained with this method are accurate for the purpose of our image reconstruction technique.

3.2. Fast Linear Image Reconstruction

A substantial decrease in the processing time can be achieved when using Equations (7), (8), (9) and (10) compared to the use of Equations (3), (4), (5) and (6). The processing time can further be reduced by calculating the inverse matrix of the regressor Q_c using the Moore-Penrose pseudoinverse [22], [27] at the time of gallery formation. In this way, the calculations at test time reduce to two matrix operations (Algorithm 2). Let $\widetilde{Q}_c$ be the pseudoinverse of

the regressor Q_c calculated at the time of gallery formation, then Equation (7) can be solved at test time as:

$$\Gamma_c = \widetilde{Q}_c X_\mu \tag{16}$$

$$\widehat{X}_c = Q_c(\widetilde{Q}_c X_\mu) \tag{17}$$

In numerical analysis theory, the least squares solution using pseudoinverse is numerically less precise than using QR decomposition. However, we did not observe any degradation in the accuracy when using the pseudoinverse. Nearly two times gain in computational efficiency was achieved by the fast linear image reconstruction for ETH-80 dataset (refer to Section 5). The gain in computational efficiency is more substantial for larger datasets.

4. Experiments and Analysis

Extensive experiments were carried out to demonstrate the performance of our technique. We evaluated our technique on three commonly used and challenging video databases, namely CMU Motion of Body dataset (CMU MoBo) [10], Youtube Celebrity dataset (YTC) [15] and Honda/UCSD dataset [17] for face recognition. ETH-80 dataset [18] was used for the task of object recognition.

We compared our technique with several prominent image set classification methods. These techniques include Face Recognition using Temporal Image Sequence (TIS) [32], Discriminant Canonical Correlation Analysis (DCC) [16], Manifold-Manifold Distance (MMD) [31], Manifold Discriminant Analysis (MDA) [29], the Linear version of the Affine Hullbased Image Set Distance (AHISD) [4], the Convex Hullbased Image Set Distance (CHISD) [4], Graph Embedding Discriminant Analysis (GEDA) [11], Sparse Approximated Nearest Points (SANP) [14], Covariance Discriminant Learning (CDL) [30], Regularized Nearest Points (RNP) [33], Mean Sequence Sparse Representation Classification (MSSRC) [21], Set to Set Distance Metric Learning (SSDML) [35] and Adaptive Deep Network Template (ADNT) [13]. We also compared our results with the Dual Linear Regression based Classifier (DLRC) [5], Multi-Manifold Deep Metric Learning (MMDML) [19], Pairwise Linear Regression Classifier (PLRC) [8] and Simultaneous Feature and Sample Reduction (SFSR) [34]. We followed the standard experimental protocols which are also followed by [13], [14], [16], [29], [30] and [31]. For comparison, we referenced the recognition results of [4], [11], [13], [14], [16], [21], [29], [30], [31], [32], [33] and [35] reported in [13]. The results of DLRC [5], MMDML [19], PLRC [8] and SFSR [34] are referenced from the respective papers. It is not feasible to compare with [19] and [34] on Youtube Celebrity Dataset because of the significantly different testing protocol used in [19] and [34]. For the results of [5] and [8] on unreported datasets, we used the implementations of these methods provided by the authors of the respective papers. We optimized the number of training and testing images in [5] and [8] for the optimal performance.

Algorithm 2: Algorithm for Fast image reconstruction and Classification

Input : Gallery image sets K_c, where $c = 1, 2, 3, ...C$. Test image set Y_μ.
Output: Class μ of test image set Y_μ.
Gallery Formation:
for c *in* 1 *to* C **do**
  for n *in* 1 *to* N **do**
    $q_c^n \in \mathbb{R}^{T\times 1} = downsample\ images\ to\ a \times b$ and $vectorize$, where $T = ab$
  end
  $Q_c \in \mathbb{R}^{T\times N} = [q_c^1 q_c^2 q_c^3 ... q_c^N]$
  $\widetilde{Q}_c = pseudoinverse(Q_c)$
end
Testing:
for m *in* 1 *to* M **do**
  $x_\mu^m \in \mathbb{R}^{T\times 1} = downsample\ images\ to\ a \times b$ and $vectorize$, where $T = ab$
end
$X_\mu \in \mathbb{R}^{T\times M} = [x_\mu^1 x_\mu^2 x_\mu^3 ... x_\mu^M]$
for c *in* 1 *to* C **do**
  $\Gamma_c = \widetilde{Q}_c X_\mu$
  $\widehat{X}_c = Q_c \Gamma_c$
  $D_c = \sqrt{\sum^T ((X_\mu - \widehat{X}_c)^2)}$
  $\Theta_c = \sum_{m=1}^{M} e^{-\alpha D_c}$
end
$\mu = arg\ \max_c(\Theta_c)$

4.1. CMU MoBo Dataset

The CMU Motion of Body Database (CMU MoBo) [10] contains videos of 25 individuals walking on a treadmill, captured from six different viewpoints. Only the videos from the front camera are used for the purpose of image set classification. All the subjects except the last one has four different videos following different walking patterns. The original purpose of this database was to advance biometric research on human gait analysis [10]. We used the video sequences of the first 24 individuals, as they contain all four walking patterns, which is similar to the previous works [5], [8], [13]. The frames of each video were considered as an image set. Similar to [4], [13], [14] and [31], we randomly selected the video of one walking pattern of each individual as the gallery image set and the other three walking patterns were considered as the test set. 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

Methods↓ \ Datasets→	MoBo	YTC	Honda	ETH-80
TIS [32]	96.81 ± 1.97	50.21 ± 3.59	88.21 ± 3.86	75.50 ± 4.83
DCC [16]	88.89 ± 2.45	51.42 ± 4.95	92.56 ± 2.25	91.75 ± 3.74
MMD [31]	92.50 ± 2.87	54.04 ± 3.69	92.05 ± 2.25	77.50 ± 5.00
MDA [29]	80.97 ± 12.28	55.11 ± 4.55	94.36 ± 3.38	77.25 ± 5.46
AHISD [4]	92.92 ± 2.12	61.49 ± 5.63	91.28 ± 1.79	78.75 ± 5.30
CHISD [4]	96.52 ± 1.18	60.42 ± 5.95	93.62 ± 1.63	79.53 ± 5.32
GEDA [11]	84.86 ± 3.24	52.48 ± 4.45	91.28 ± 5.82	79.50 ± 5.24
SANP [14]	97.64 ± 0.94	65.60 ± 5.57	95.13 ± 3.07	77.75 ± 7.31
CDL [30]	90.00 ± 4.38	56.38 ± 5.31	98.97 ± 1.32	77.75 ± 4.16
RNP [33]	96.11 ± 1.43	65.82 ± 5.39	95.90 ± 2.16	81.00 ± 3.16
MSSRC [21]	97.50 ± 0.88	59.36 ± 5.70	97.95 ± 2.65	90.50 ± 3.07
SSDML [35]	95.14 ± 2.20	66.24 ± 5.21	86.41 ± 3.64	81.00 ± 6.58
DLRC [5]	91.60 ± 2.78	65.55 ± 5.16	92.31*	86.5 ± 6.0323
MMDML [19]	97.8 ± 1.0	—	**100.00 ± 0.0**	94.5 ± 3.5
ADNT [13]	**97.92 ± 0.73**	**71.35 ± 4.83**	**100.00 ± 0.0**	**98.12 ± 1.69**
PLRC [8]	93.74 ± 4.3	61.28 ± 6.37	89.74*	87.72 ± 5.67
SFSR [34]	96.0*	—	96.8*	—
Ours	**98.33 ± 1.27**	**66.45 ± 5.07**	**100.00 ± 0.0**	**94.75 ± 4.32**

Table 1. Average classification accuracies and standard deviations on CMU MoBo (MoBo) [10], YouTube Celebrity (YTC) [15], UCSD/Honda (Honda) [17] and ETH-80 [18] datasets. Both algorithms for the proposed technique have the same accuracy.
* Indicates use of different experimental protocol as authors have reported results for only one fold experiments.

lower than the number of pixels. We randomly selected a small number of frames i.e., 50 from each gallery video. The face from each frame was automatically detected using the Viola and Jones face detection algorithm [28]. Similar to [13], the images were resampled to the resolution of 40×40 and converted to grayscale. Histogram equalization was applied to increase the contrast of images. Different to [5], [8], [13], we did not use any LBP features, and performed experiments on raw images. We used $\alpha = 0.2$ in Equation (12). The experiments were repeated for 10 times with different random selections for 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. We achieved the best classification accuracy on MoBo dataset among all compared techniques. Table 1 provides the average accuracy of our technique along with a comparison with other methods.

4.2. YouTube Celebrity Dataset

The Youtube Celebrity (YTC) Dataset [15] contains 1910 video clips of 47 celebrities and politicians. This is the largest dataset used for image set classification. These noisy real world videos, downloaded from YouTube, have low resolution and are recorded at high compression rates. The Viola and Jones algorithm [28] failed to detect faces for a large number of frames. Therefore, similar to [13], the Incremental Learning Tracker [24] was used to track the faces in video clips. Although the cropped face region was not uniform across frames, we decided to use the automatically tracked faces without any refinement. As proposed in [13], [14], [29], [30], [31], five fold cross validation was used for experiments. The datset was divided into five folds while minimizing the overlap between the various folds. Each fold contains 423 video clips with 9 video clips per individual. Out of 9 video clips per individual, three videos were randomly selected as the gallery set while the remaining six were used as six separate test image sets. All the tracked face images were resampled to the resolution of 30×30 and converted to grayscale, following the protocol of [13]. Histogram equalization was applied to enhance the contrast of images. For the gallery image set we randomly selected 20 images from each of the three gallery videos per individual per fold. If any gallery video clip had less than 20 frames, all the images of that video were used for gallery formation. In this way each gallery set had a maximum of 60 images. We used $\alpha = 10.5$ in Equation (12). Our technique achieved the highest accuracy among all the parametric and non-parametric methods. Deep Learning based ADNT [13] has a better classification accuracy, however, it should be noted that our method uses significantly less training data compared to [13] and is much faster than [13] (refer to Section 5). Moreover our technique does not require any parameter fine tuning or training which makes it more suitable for real life applications. Table 1 summarizes the average accuracies of the different techniques on YouTube Celebrity dataset.

4.3. UCSD/Honda Dataset

The UCSD/Honda Dataset [17] consists of 59 videos of 20 individuals. The number of videos for each individual varies from one to five. The database was originally developed to provide a standard video database to evalu-

Datasets ↓	Methods ↓	Resolution used by [13]	20 × 20 Resolution	15 × 15 Resolution
MoBo	ADNT [13]	97.92 ± 0.73	91.81 ± 2.40	90.56 ± 3.13
	Ours	**98.33 ± 1.27**	**98.75 ± 1.38**	**99.31 ± 1.18**
YTC	ADNT [13]	**71.35 ± 4.83**	61.06 ± 5.67	57.66 ± 4.85
	Ours	66.45 ± 5.07	**64.40 ± 5.22**	**65.25 ± 5.05**
Honda	ADNT [13]	100.00 ± 0.00	100.00 ± 0.00	99.74 ± 0.81
	Ours	100.00 ± 0.00	100.00 ± 0.00	**100.00 ± 0.00**
ETH-80	ADNT [13]	**98.12 ± 1.69**	88.75 ± 6.26	90.25 ± 4.63
	Ours	94.75 ± 4.32	**95.50 ± 4.04**	**92.75 ± 6.39***

Table 2. Average classification accuracies and standard deviations on the low resolutions of our technique compared with ADNT [13].
* The slight decrease in performance is due to the nearly equal number of gallery images and the number of pixels (refer to Section 3 for details).

ate face tracking and recognition algorithms [17]. All the videos contain significant head rotations and pose variations. Moreover some of the video sequences also contain partial occlusions in some frames. We followed the same experimental protocol as [13], [14], [17], [29] and [31]. The face from each frame of videos was automatically detected using Viola and Jones face detection algorithm [28]. Similar to [13], the detected faces were downsampled to the resolution of 20×20 and converted to grayscale. Histogram equalization was applied to increase the contrast of images. The images were standardized by subtracting the mean and dividing by the standard deviation. We randomly selected one video from each of the 20 individuals as the gallery image set while the remaining 39 videos were used as the test image sets. In order to keep the number of gallery images considerably less than the number of pixels (refer to section 3), we randomly selected a small number of frames i.e., 50 from each gallery video. We used $\alpha = 0.2$ in Equation (12). To improve the consistency in scores we repeated the experiment 10 times with different random selections of gallery images, gallery image sets and test image sets. Our technique achieved a perfect classification accuracy while using a significantly less number of gallery images. Table 1 summarizes the average identification rates of our technique compared to other image set classification techniques.

4.4. ETH-80 Dataset

The ETH-80 dataset [18] consists of eight object categories and each object category has ten different image sets. Each image set consists of 41 images of the object taken from different view angles. The cropped images containing only the object without any border area were used. The images were resized to the resolution of 32×32 to follow the protocol of [13]. The images were converted to grayscale and were standardized by subtracting the mean and dividing by the standard deviation. Similar to [13], [16], [29] and [30], five image sets of each object category are randomly selected as the gallery set while the other five are considered to be independent test image sets. We used $\alpha = 0.2$ in Equation (12). We repeated the experiments 10 times for different random selections of gallery and test sets. The performance of our technique is comparable to the state of the art deep learning technique [13]. Table 1 summarizes the results of our technique compared to other methods.

4.5. Experiments at low resolution

We carried out further experiments at lower resolutions to demonstrate the efficacy of our technique. We also compared the performance of our technique with ADNT [13] using low resolution data. We kept all other experimental settings and protocols the same as in the previous sections. For [13], we used the implementation provided by the authors and kept all parameters settings the same as recommended in their paper. Table 2 shows the average classification accuracies which demonstrates the superior performance of our technique at low resolution. On MoBo dataset, the performance improves at lower resolution. The change in the performance of our technique for Youtube Celebrity dataset is negligible as compared to the degradation in performance of ADNT [13]. For Honda dataset, the classification accuracy remains a perfect score with the change in resolution. On ETH-80 Dataset, we achieved the best performance at 20×20 resolution. This is due to the fact that at 15×15 resolution, the number of gallery images is nearly equal to number of pixels (refer to Section 3 for details). By reducing the number of gallery images at 15×15 resolution, we achieved an average classification accuracy of 95.25%. Overall, there is no significant change in the classification accuracy of our technique with the change in resolution. Compared to ADNT [13], our technique consistently achieved better performance using low resolution data. This shows that our technique is more suitable to applications where the data is of very low resolution e.g., CCTV surveillance.

5. Computational Time Analysis

The proposed technique achieves the fastest timing performance as compared to other techniques. Table 3 shows the training time for various methods and the test time required to classify an image set on the ETH-80 dataset using a modern CPU with 8 GB RAM. The proposed technique requires no training. Although our technique reconstructs each image of the test image set from all the gallery image

Methods ↓	Total Training Time (seconds)	Test Time per image set (seconds)
TIS [32]	NR	0.045
DCC [16]	13.36	0.311
MMD [31]	NR	8.43
MDA [29]	**1.22**	**0.005**
AHISD [4]	NR	0.095
CHISD [4]	NR	0.213
GEDA [11]	2.7	0.068
SANP [14]	NR	105.7
CDL [30]	76.21	1.40
RNP [33]	NR	0.027
MSSRC [21]	NR	4.78
SSDML [35]	21.92	0.577
ADNT [13]	**278.8**	**0.026**
Ours	**NR**	**0.0046**
Ours (Fast)	**NR**	**0.0028**

Table 3. Computational Time Analysis on ETH-80 dataset. NR shows that the method does not require training.

sets, but due to the efficient matrix representation (refer to Section 3 and Section 3.2), we achieved timing efficiency superior to the other methods.

6. Discussion

The techniques of [5] and [8] are based on linear models. However, our technique is remarkably different from [5], [8]. The works of [5] and [8] consider test image set as a subspace of a high dimensional space and use the distance between the test and training image sets to determine the class of the test image set. On the contrary, we treat each image in the test image set independently and consider them as points in a high dimensional space. To determine the distance between subspaces, [5] estimates a virtual image by using the last image of each image set along with the variations between the training and test sets [8]. The distance of image sets from the virtual image is used as the distance metric. The work of [8] is an extension of [5] where instead of the last image of each image set, the mean image is used along with the concept of related and unrelated subspaces. On the other hand, we reconstruct each test image from the gallery subspaces and use weighted voting with the Euclidean distances between the original and reconstructed test images. Weighted voting increases the robustness of our system to any noise and outliers in the test image set. Both [5] and [8] can only work for small test sets due to the limitation that the combined number of images in the test and the training image sets should be much less than the number of features in the feature vectors. In addition to the gallery image sets, they also use test image sets as regressors, which render them prone to the problem of rank deficient matrix at test time. Our technique does not impose any constraints on the number of images in the test set and can work with any number of test images. Moreover, we performed all of our experiments on raw images to demonstrate the generalizability of our technique. In contrast to [5], [8], once any singularity is removed in the regressor Q_c at the time of the gallery formation (refer to Section 3.1), our technique is immune to the problem of rank deficient matrix at test time. Our technique can process whole image sets simultaneously and also has the capability to process one image at a time and update the class decision in real time which makes it suitable for live video surveillance (e.g., CCTV).

The accuracy of the proposed technique is superior to all parametric and non-parametric methods. The deep learning technique ADNT [13] has a better accuracy on the Youtube Celebrity dataset and the ETH-80 dataset at high resolutions. However, ADNT needs a lot of training data and relies on handcrafted LBP features. ADNT uses Restricted Boltzman Machine for parameter initialization and requires a lot of fine tuning. On the other hand, our technique uses only a fraction of the training data and achieves comparable results, using only the raw images. Moreover, our technique has produced superior results at lower resolutions, compared to ADNT, and is ten times faster than ADNT at test time. Our technique can also be easily generalized to new data. The capability to work with less training data and at low resolution deems our technique suitable for scenarios where only scarce training data is available and where fast decisions are required.

7. Conclusion

In this work, a novel image set classification technique is proposed. The proposed technique uses linear regression to reconstruct images of the test image set from gallery image sets and uses the accumulative weighted reconstruction error to decide for the class of the test image set. The technique requires less training data compared to other image set classification methods and can work effectively at very low resolution. Extensive experimental analysis has been presented on a number of popular and challenging datasets to demonstrate the superior performance of our technique. Through the efficient matrix implementation, the proposed technique achieves the fastest performance time. The technique can easily be scaled from processing one frame at a time (for live video acquisition) to processing all of the test data at once (for faster performance). All these factors make our technique ideal for image set classification applications.

Acknowledgment

This work is supported by SIRF Scholarship from the University of Western Australia (UWA) and Australian Research Council (ARC) grant DP150100294

References

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

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

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[4] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2567–2573. IEEE, 2010.

[5] L. Chen. Dual linear regression based classification for face cluster recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2673–2680. IEEE, 2014.

[6] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *European Conference on Computer Vision*, pages 766–779. Springer, 2012.

[7] Y.-C. Chen, V. M. Patel, S. Shekhar, R. Chellappa, and P. J. Phillips. Video-based face recognition via joint sparse representation. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.

[8] Q. Feng, Y. Zhou, and R. Lan. Pairwise linear regression classification for image set retrieval. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4865–4872, 2016.

[9] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[10] R. Gross and J. Shi. The cmu motion of body (mobo) database. 2001.

[11] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2705–2712. IEEE, 2011.

[12] M. Hayat, M. Bennamoun, and S. An. Learning non-linear reconstruction models for image set classification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1907–1914, 2014.

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

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

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

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

[17] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *2003 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–313. IEEE, 2003.

[18] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *2003 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–409. IEEE, 2003.

[19] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1137–1145, 2015.

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

[21] E. G. Ortiz, A. Wright, and M. Shah. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3531–3538, 2013.

[22] R. Penrose. On best approximate solutions of linear matrix equations. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 52, pages 17–19. Cambridge Univ Press, 1956.

[23] W. H. Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[24] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, 2008.

[25] T. P. Ryan. *Modern regression methods*, volume 655. John Wiley & Sons, 2008.

[26] G. A. Seber and A. J. Lee. *Linear regression analysis*, volume 936. John Wiley & Sons, 2012.

[27] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*, volume 12. Springer Science & Business Media, 2013.

[28] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[29] R. Wang and X. Chen. Manifold discriminant analysis. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 429–436. IEEE, 2009.

[30] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2496–2503. IEEE, 2012.

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

[32] O. Yamaguchi, K. Fukui, and K.-i. Maeda. Face recognition using temporal image sequence. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 318–323. IEEE, 1998.

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

[34] M. Zhang, R. He, D. Cao, Z. Sun, and T. Tan. Simultaneous feature and sample reduction for image-set classification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1401–1407. AAAI Press, 2016.

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