Simultaneous Feature and Sample Reduction for Image-set Classification

Man Zhang, Ran He*, Dong Cao, Zhenan Sun, Tieniu Tan
Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition,
Institute of Automation, CAS Center for Excellence in Brain Science and Intelligence Technology,
Chinese Academy of Sciences, Beijing 100190, China
{zhangman, rhe, dong.cao, znsun, tnt}@nlpr.ia.ac.cn

Abstract

Image-set classification is the assignment of a label to a given image set. In real-life scenarios such as surveillance videos, each image set often contains much redundancy in terms of features and samples. This paper introduces a joint learning method for image-set classification that simultaneously learns compact binary codes and removes redundant samples. The joint objective function of our model mainly includes two parts. The first part seeks a hashing function to generate binary codes that have larger inter-class and smaller intra-class distances. The second one reduces redundant samples with discrete constraints in a low-rank way. A kernel method based on anchor points is further used to reduce sample variations. The proposed discrete objective function is simplified to a series of sub-problems that admit an analytical solution, resulting in a high-quality discrete solution with a low computational cost. Experiments on three commonly used image-set datasets show that the proposed method for the tasks of face recognition from image sets is efficient and effective.

1 Introduction

Image-set classification generally aims to recognize an object from a set of related images that form an image set (Kim, Arandjelovic, and Cipolla 2007). It is the assignment of a classification label to a given image set. With the development of imaging technology, it has been widely applied in various real-world scenarios such as recognition from surveillance videos, camera networks and personal albums. Compared with single image based classification, image-set classification often offers more promises because it can effectively represent image appearance variations incurred by pose, illumination, expression, and etc (Hayat, Bennamoun, and An 2015).

In the last decade, image-set classification has drawn significant attention from the research community and many methods have been developed. Most of these methods can be categorized into two groups: finding a representation of an image set and defining a distance metric between two image sets. In these methods, an image set has been modeled by a subspace (Kim, Arandjelovic, and Cipolla 2007)(He et al. 2015), a combination of subspaces (Nishiyama et al. 2007), a combination of dictionaries (Chen et al. b)(Lu et al. 2014)(Zhang, He, and Davis 2014), sparse representation (Ortiz, Wright, and Shah 2013), Grassmann Manifold (Harandi et al. 2010)(Huang et al. 2015b), Riemannian manifold (Wang et al. 2015)(Huang et al. 2015a). Recently, (Hayat, Bennamoun, and An 2015) proposed a deep learning based representation model whose parameters are initialized via Gaussian Restricted Boltzmann Machines. In addition, nearest points (Hu, Mian, and Owens 2012)(Yang et al. 2013) were introduced to calculate the between-set distance.

The existing image-set classification methods often perform classification tasks based on hundreds of floating point features, and store almost every sample from each image set. When image-set classification is applied to surveillance videos or camera networks, dense features and large-scale registered samples result in tremendously large time and space complexity because there can be (many) thousands of image samples in a video clip. Dense features and large-scale samples have become a computational bottleneck of real-life applications, resulting in two open issues. The first issue is to learn compact binary codes to represent an image set. The second one is to reduce the number of large-scale and similar samples in an image set.

To address the two issues simultaneously, we present a novel method to jointly learn compact binary codes and reduce the number of different samples in each image set. Firstly, we formulate the proposed method as a joint learning problem that includes two parts. The first part seeks a hashing function to generate binary codes that have larger inter-class distances and smaller intra-class distances. The second reduces redundant samples with discrete constraints in a low-rank way; similar samples from the same image set are forced to have the same binary codes. Secondly, we relax the proposed discrete objective function to a series of sub-problems to admit an analytical solution. Then we can obtain a high-quality discrete solution with a very low computational cost. A kernel method based on anchor points is further used to handle complex and nonlinear variations of images, which allows implicit and nonlinear feature mappings. Finally, given a hashing function and a new image set (not in the training set), we formulate the binarization of an image set as a structured learning problem with discrete constraints. Experimental results on two standard datasets and one large-scale dataset (the number of comparisons is more than 900 million) show that this structured formulation can significantly reduce the number of different samples and also demonstrate the efficiency and effectiveness of the proposed method for the tasks of face recognition from image sets.

The major contributions of this work lie in three-folds:

- In our work, a joint learning framework is proposed for image-set classification that reduces the redundancy of both features and samples at the same time. To the best of our knowledge, the proposed method is the first one to address the issue of learning hashing function for image sets.
- We treat the high correlation between samples (especially for videos) as an output structure of binary codes, and formulate this kind of correlation as a matrix trace term (low-rank) with discrete constraints. This formulation makes the proposed framework more flexible for image-set based problems and results in high quality discrete codes.
- Different from other single image based hashing methods that employ the sign function to binarize each sample independently, we treat the samples in an image set as a whole one and propose a structured method to binarize this set of images. The binarization depends on not only the hashing function but also the binary values of similar samples.

2 Simultaneous Feature and Sample Reduction

2.1 Motivation

In image-set classification, each image set often consists of a large number of images and each image is represented by a high-dimensional and dense feature. To reduce time and space complexity, we aim to learn compact binary codes for image-set classification. At the same time, since the images (especially video images) from the same image set are often similar, we expect that similar images have the same (or similar) binary codes.

Consider a training set X from C classes, which consists of n image samples x_i $(1 \leq i \leq n)$ in a high dimensional Euclidean space R^d. We want to seek a binary representation for all images in X, which forms a binary matrix $B = [B_1, \ldots, B_n] \in R^{m \times n}$. Then the problem of learning binary codes can be formulated as the following linear regression problems with discrete constraints,

$$\begin{aligned} &\min_{W,B} \left\|W^T X - B\right\|_F^2 + \lambda_1 \left\|W\right\|_F^2, \\ &s.t.\ B_{ij} \in \{-1, 1\} \end{aligned} \tag{1}$$

where λ_1 is constant and $\|.\|_F$ indicates matrix Frobenius norm. In hashing (Wang et al. 2014), projection matrix $W \in R^{m \times d}$ defines a linear hashing function for data X.

Since the image appearance variations in an image set are often large, a linear hashing function may not handle all variations very well. Inspired by the supervised hashing methods with kernels(Liu et al. 2012)(Shen et al. 2015), we can map data X into a kernel space spanned by k randomly selected anchor points $\{a_j\}_{j=1}^k$. Then (1) becomes,

$$\begin{aligned} &\min_{W,B} \left\|W^T K - B\right\|_F^2 + \lambda_1 \left\|W\right\|_F^2, \\ &s.t.\ B_{ij} \in \{-1, 1\} \end{aligned} \tag{2}$$

where $K = [K_1, \ldots, K_n] \in R^{k \times n}$. Each K_i is obtained by the RBF kernel mapping:

$$K_i = [\exp(\|x_i - a_1\|_2^2 / \sigma), \ldots, \exp(\|x_i - a_k\|_2^2 / \sigma)],$$

where σ is the kernel width. Since the binary samples in B are from C classes, we expect that these samples have larger inter-class and smaller intra-class distances, resulting in the following criteria,

$$\Omega(B) = \tfrac{1}{2} \sum_{m,n \in c} h(B_m, B_n) - \tfrac{1}{2} \sum_{m \in c} \sum_{n \in c'} h(B_m, B_n) \tag{3}$$

where $c \in \{1 : C\}$, $c \neq c'$ and h is a distance in the hamming space.

Considering that the samples from an image set can be correlated, we further introduce a low-rank constraint to (2) to encourage B^c from the c-th class to be correlated. This constraint reduces the redundancy of samples and potentially corrects some binary codes whose corresponding values $(W^T K)$ are close to the separating hyperplane (i.e., zero). Combining all points together, we have the following hashing problem for image-set classification,

$$\begin{aligned} &\min_{W,B} \left\|W^T K - B\right\|_F^2 + \Omega(B) + \sum_c \|B^c\|_* + \lambda_1 \|W\|_2, \\ &s.t.\ B_{ij} \in \{-1, 1\} \end{aligned} \tag{4}$$

where $\|.\|_*$ denotes the matrix trace norm (i.e., the sum of its singular values) and B^c contains the columns in B that belong to the c-th class. When we apply the values $\{-1, 1\}$ to indicate binary values $\{0, 1\}$, the low-rank constraint in (4) might lead to a special case. That is, let codes $B_1 = (1, 1, 1, 1)^T$, $B_2 = (-1, -1, -1, -1)^T$, and then matrix $B = [B_1, B_2]$ is of rank 1. The matrix B has the lowest rank, but the distance between B_1 and B_2 is the largest. Fortunately, we have another constraint in (3) to avoid this extreme case. Since (4) involves a non-convex trace norm and discrete constraints, it is a difficult optimization problem.

2.2 Solution

Directly minimizing the non-convex problem in (4) is quite hard. Hence we first introduce an auxiliary variable $S \in R^{m \times n}$ to relax (4), resulting in the following minimization problem,

$$\begin{aligned} &\min_{W,B,S} \left\|W^T K - B\right\|_F^2 + \Omega(S) + \sum_c \|B^c\|_* + \lambda_1 \|W\|_2, \\ &s.t.\quad S = B,\ B_{ij} \in \{-1, 1\} \end{aligned} \tag{5}$$

To go down the objective function in (5), we employ an iterative block coordinate descent method. Then we have the following sub-problems,

$$\min_{W} \quad \left\|W^TK - B^{t-1}\right\|_F^2 + \lambda_1 \left\|W\right\|_2, \tag{6}$$

$$\min_{S} \quad \Omega(S|B^{t-1}), \tag{7}$$

$$\min_{B^c} \quad \left\|A^{ct} - B^c\right\|_F^2 + \left\|B^c\right\|_*, \tag{8}$$

$$s.t. \quad B = S^t,\ B_{ij} \in \{-1, 1\}$$

where $A^t = (W^t)^TK$, A^{ct} contains the columns in A^t that belong to the c-th class, and $\Omega(S|B^{t-1})$ indicates that we seek the solution of $\Omega(S)$ when B^{t-1} is given. That is, we only solve one sub-problem at a time by fixing other variables.

By setting the derivative of (6) equal to zero, we have the analytic solution of (6) as follows,

$$W^t = (KK^T + \lambda_1 I)^{-1}K\{B^{(t-1)}\}^T, \tag{9}$$

where I is an identity matrix. Given B^{t-1}, we can apply sub-gradient descent method to solve the sub-problem in (7). Since S and B^{t-1} are all binary codes, (7) can be minimized very efficiently. Here, we adopt the implementation provided by (Rastegari, Farhadi, and Forsyth 2012) to obtain a local solution of S.

By substituting the first equality constraint into sub-problem (8), we can reformulate (8) as follows,

$$\min_{B^c} \left\|A^{ct} - B^c\right\|_F^2 + \lambda_2 \left\|B^c - S^{ct}\right\|_F^2 + \left\|B^c\right\|_*, \tag{10}$$
$$s.t. \quad B_{ij} \in \{-1, 1\}$$

To minimize the low-rank problem in (10), we first need to introduce a variational formulation for the trace norm (Grave, Obozinski, and Bach 2011),

Lemma 1 *Let $B \in R^{m\times n}$. The trace norm of B is equal to:*

$$\|B\|_* = \tfrac{1}{2} \inf_{L\geq 0} tr\left(B^TL^{-1}B\right) + tr(L) \tag{11}$$

and the infimum is attained for $L = (BB^T)^{1/2}$.

Using this lemma, we can reformulate (10) as,

$$\min_{B^c} \min_{L\geq 0} \left\|A^{ct} - B^c\right\|_F^2 + tr(B^{cT}L^{-1}B^c) \tag{12}$$
$$+\lambda_2 \left\|B^c - S^{ct}\right\|_F^2 + tr(L)\ s.t.\ B^c_{ij} \in \{-1, 1\}$$

The problem in (12) can be alternately minimized (Grave, Obozinski, and Bach 2011)(Wang et al. 2013)(He, Tan, and Wang 2014). Fixing L, we can employ the discrete cyclic coordinate descent method to obtain B^c. For simplicity, we make use of a simple way to compute B^c. That is, disregarding the integer constraint and setting the derivative of (12) equal to zero, the solution of B^c has the following form,

$$B^c = ((1 + \lambda_2)I + L^{-1})\backslash(A^{ct} + \lambda_2 S^{ct}). \tag{13}$$

Given a floating point B^c in one iteration, we can use the sign function $sgn(.)$ to obtain binary value $sgn(B^c)$. Experimental results show that the learned binary codes are good enough for image-set classification.

Algorithm 1: Simultaneous Feature and Sample Reduction (SFSR)

```
Input: Data matrix $X \in R^{d\times n}$
Output: Hashing function $W \in R^{d\times m}$
Normalize each $x_i$ to unit $\ell_2$-norm
Compute kernel matrix $K$ by randomly selecting $k$ samples from $X$ and centralize $K$ to have zero mean
Initialize $B^0$ by PCA+ITQ
repeat
    Compute $W^t$ via (6)
    Compute $S^t$ via (7)
    Compute $B^t$ via (8)
    t ← t+1
until The variation of $B$ is smaller than a threshold
```

Algorithm 1 summarizes the procedure to simultaneously reduce features and samples for image-set classification. In Algorithm 1, we have to give an initial value of B^0. Fortunately, error correcting code (Kittler et al. 2001) and information theoretic rules (Chen et al. a) provide an efficient way to initialize B^0. In error correcting code methods, the binary codes for one class are often unique. Here, we employ unsupervised version of iterative quantization (PCA+ITQ) (Gong and Lazebnik 2011) to obtain the binary code of the mean image of one class. Then, these binary codes are used as B^0. Experimental results in Section 3.2 show that Algorithm 1 is not very sensitive to the setting of B^0. The threshold of the variation of B can be determined by cross-validation on the training set. When the number of iterations is large enough, Algorithm 1 is not sensitive to the setting of this threshold. Throughout this paper, λ_1 and λ_2, training and selected based on the datasets, are empirically set to 1 and 0.1, respectively.

2.3 Image-set Classification

Given a learnt hashing function $f_W^H(.)$ and a new probe dataset K^p, one still needs to quantize floating-point $f_W^H(K^p)$ to discrete values (Kong and Li 2012). The existing hashing methods often assume that the samples in K^p are independent and resort to the sign function $sgn(.)$ to obtain binary value as follows,

$$B_{ij} = sgn(f_W^H(K_{ij}^p)) \tag{14}$$

For single-image based hashing methods, we can further formulate the quantization as a discrete optimization problem as follows,

$$\min_{B_i} \quad \left\{\left\|E_i\right\|_2^2 + \left\|B_i\right\|\right\}, \tag{15}$$

$$s.t. \quad B_i = f_W^H(K_i^p) + E_i,\ B_i \in \{-1, 1\}$$

(a) the Honda/UCSD dataset (b) the CMU Mobo dataset (c) the YouTube Celebrities dataset

Figure 1: Image samples of three different classes in the three image-set databases respectively. There are large appearance variations incurred by pose, illumination and expression.

where $\|.\|$ is a potential norm from the prior knowledge of data.

Having considered that the samples in the image set K^p are often correlated, we introduce the following low-rank problem with discrete constraints for image-set classification,

$$\begin{aligned} \min_{B} \quad & \left\{ \|E\|_F^2 + \|B\|_* \right\} \\ s.t. \quad & B = f_W^H(K^p) + E, \; B_{ij} \in \{-1, 1\} \end{aligned} \tag{16}$$

Since the low-rank constraint in (16) tends to make the column samples in B correlated, it also tends to reduce the number of different samples in B. By substituting the first constraint into (16), (16) becomes,

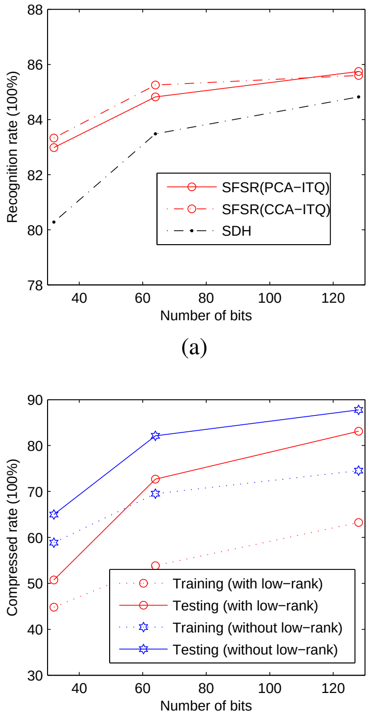
$$\begin{aligned} \min_{B} \quad & \left\{ \left\|B - f_W^H(K^p)\right\|_F^2 + \|B\|_* \right\} \\ s.t. \quad & B_{ij} \in \{-1, 1\} \end{aligned} \tag{17}$$

Comparing (17) with (8), we can employ the same minimization method for (8) to minimize (17). Experimental results in Section 3.2 demonstrate that the usage of (16) to quantize floating-point features can significantly reduce the number of different samples.

Since B is a binary value matrix and has a small number of different samples, we can make use of a simple yet efficient voting method for image-set classification. That is, a simple nearest neighbor classifier for each unique code in B with voting is used as classifier to report recognition rates. The class label of the majority class in an image set is taken as the final class label of this set.

3 Experiments

We evaluate the performance of the proposed method for the tasks of face recognition from image sets. Since there seem no existing hashing methods that are particularly designed for image-set classification, we compare the proposed method in several different settings with state-of-the-art hashing methods. Meanwhile, we also compare our method with state-of-the-art image-set classification techniques. Experiments are performed on two standard image-set classification datasets, i.e., the Honda/UCSD dataset (Lee et al. 2003), the CMU Mobo Dataset (Gross and Shi 2001), and one large-scale dataset (the number of comparisons is larger than 900 million), i.e., the YouTube Celebrities Dataset (Kim et al. 2008).



(a)

(b)

Figure 2: Performance of the proposed SFSR in several different settings. (a) Recognition rates with different initializations on the Youtube dataset. (b) Compressed rates with/without using (16) on the training set and testing set of the Honda dataset.

3.1 Evaluation Protocols

To systematically evaluate different methods, we make use of recognition rates and compressed rates as two evaluation protocols, which measure the performance of SFSR in terms of feature and sample reduction, respectively. Recognition rate is defined as the ratio between the number of correctly classified image set and the total number of image sets. A higher recognition rate of an algorithm indicates that the algorithm correctly classifies more image sets. Compressed rate is the compressed ratio of samples during image-set classification, i.e., compressed rate = the number of unique samples/the total number of samples. A lower compressed ratio of an algorithm indicates that the algorithm needs less storage space (and as a consequence less computational time).

3.2 Algorithmic Analysis

In this subsection, we give a deep analysis of the proposed method in several different settings. In Algorithm 1, we should give an initial value to B^0. We further study of the robustness of Algorithm 1 to B^0 by using different initializations. Unsupervised version (PCA+ITQ) and supervised version (CCA+ITQ) of ITQ (Gong and Lazebnik 2011) are employed to initialize B^0. Fig. 2 (a) shows the recognition rates of SFSR by using these two initializations on the YouTube celebrities dataset. From the results, we observe that there are only small variations between two initializations. That is, Algorithm 1 is not very sensitive to the initialized value of B^0.

Low-rank constraints play an important role in the proposed methods during both training and testing. Hence we further study the performance of SFSR with/without using low-rank constraints. A simple way to testify performance by whether using (16) to obtain binary codes or not. If (16) is not used, one can directly obtain binary codes via $sgn(f_W^H(K^p))$. Fig. 2 (b) shows the compressed rates of SFSR with/without using (16) on the Honda dataset. We observe that compressed rates are significantly reduced when (16) is used to obtain binary codes. This indicates that the low-rank constraints in SFSR can effectively and efficiently model the correlation relationship between the images in an image set and are applicable for image-set classification problems. Moreover, $\Omega(B)$ reduces intra-class variation so that some extreme cases (e.g., $B_1 = (1,1,1,1)^T, B_2 = (-1,-1,-1,-1)^T$) do not occur. We aim to find binary codes (0,1) and can impose low-rank constraint on 0-1 values to avoid the extreme cases.

3.3 Numerical Results

The Honda/UCSD Dataset consists of 59 video sequences for 20 persons. Each person has at least 2 videos that contain large pose and expression variations. The numbers of the images in these sequences vary from 12 to 645. Fig. 1 (a) shows some cropped images from this dataset. We make use of the standard training/testing protocol in (Wang et al. 2008)(Zhang, He, and Davis 2014): 20 sequences are used for training and the remaining 39 sequences for testing. All the video frames are used to report classification results.

Fig. 3 (a) and Fig. 4 (a) show recognition rates and compress rates on the Honda database, respectively. We compare the proposed SFSR with several other state-of-the-art hashing methods. SFSR can reach the best recognition result in different numbers of bits in Fig. 3 (a). Meanwhile, the compressed rate of SFSR is better than most of the compared algorithms, similar to FastHash and lower than SDH. Since the binary codes of SDH are constrained to form a basis of sparse label matrix, SDH potentially reduces the number of different samples in an image set. However, this constraint in SDH may make the binary codes of SDH overfit on the training set so that SDH can not achieve a higher recognition rate.

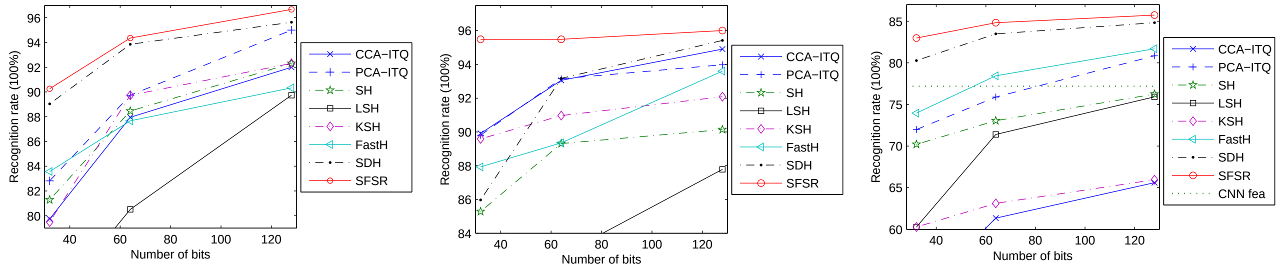
The CMU Mobo Dataset was originally published for human pose identification. It contains 96 sequences of 24 different subjects, which are captured in four different walking patterns (slow, fast, inclined, and carrying a ball) by using multiple cameras. Fig. 1 (b) shows some cropped images from three subjects. We follow the standard training/testing configuration in (Wang et al. 2008)(Zhang, He, and Davis 2014). One video was randomly chosen as training and the remaining three for testing.

Fig. 3 (b) and Fig. 4 (b) illustrate the experimental results on the Mobo database. SFSR is robust to appearance variations in videos and extract discriminative binary codes, thus SFSR also achieve the highest recognition rate in this data set. When the number of bits is small, the compressed rate of SFSR is higher than that of SDH. However, the compressed rate of SFSR is the lowest when the number of bits grows. The proposed method can remove redundancy and represent the useful image information effectively and efficiently.

The YouTube Celebrities Dataset is composed of 1910 video clips of 47 human persons (actors, actresses, and politicians) from the YouTube website. Roughly 41 clips were segmented from 3 unique videos for each person, and are low resolution and highly compressed. Fig. 1 (c) shows some cropped facial images whose sizes are 30 × 30. This dataset is challenging because it contains many tracking errors in cropped faces and large appearance variations (e.g., pose, illumination and expressions). Following the standard protocol, the testing dataset is composed of 6 test clips, 2 from each unique video, per person. The remaining clips were used as the input to the CNN to learn a 1152-D feature representation. One frame of video (one single image) is fed into the CNN at a time. We randomly selected 3 training clips, 1 from each unique video. The number of images in the testing set is 44,172. The average number of training images for hashing algorithm is larger than 20,000. Hence, the average number of image comparisons is more than 900 million.

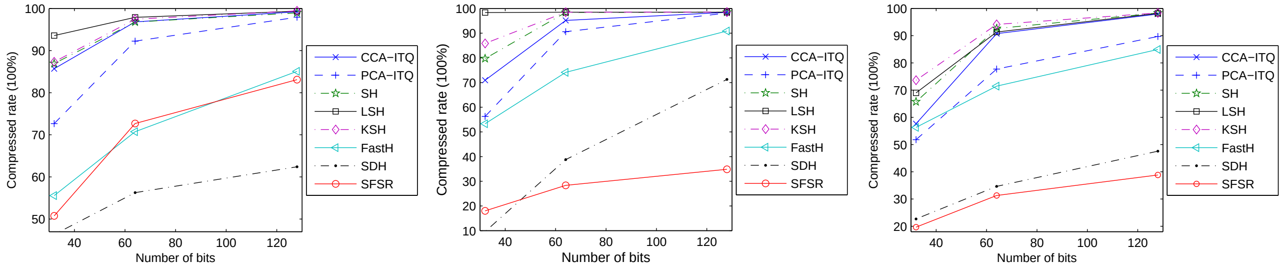
The experimental results on the Youtube database are shown in Fig. 3 (c) and Fig. 4 (c). Although CNN features are used, the recognition rates of all methods on this dataset are significantly lower than those on the previous two datasets. This is because there are many tracking and detection errors (as shown in Fig. 1 (c)) so that cropped face images are not well alignment. Face alignment is a key step for face recognition. We observe that SFSR obtains the best recognition rates and the lowest compressed rates in each number of bits, and further improve the recognition performance of CNN features. SFSR with discrete constraints is flexible for image-set classification and results in high compressed and discriminative codes. Thus, a small number of bits can be selected to decrease computation complexity and time consuming without decreasing the recognition performance in the database.

Compared with Image-set Classification Methods In this subsection, we compare the proposed method with state-



(a) Honda (b) Mobo (c) Youtube

Figure 3: Recognition rates of different binary code learning methods on the three image-set classification databases. A higher recognition rate of an algorithm indicates that the algorithm correctly classifies more image sets.



(a) Honda (b) Mobo (c) Youtube

Figure 4: Compressed rates of different binary code learning methods on the three image-set classification databases. A lower compressed ratio of an algorithm indicates that the algorithm needs smaller storage space and computational time.

of-the-art image-set classification methods, including, regularized nearest points (RNP) (Yang et al. 2013), mean sequence sparse representation-based classification (MSSRC) (Ortiz, Wright, and Shah 2013), single Gaussian model (SGM) (Huang et al. 2015a), projection metric learning (PML) (Huang et al. 2015b), discriminant analysis on Riemannian Manifold (DARG) (2015), and Deep Reconstruction Model (DRM) (2015). As in (Yang et al. 2013)(Ortiz, Wright, and Shah 2013)(Zhang, He, and Davis 2014) , we directly cited the best recognition rates of these methods from the literature.

Table 1 tabulates the classification accuracy of each image set method. It is shown that on the large-scale dataset the proposed method significantly outperforms all other approaches. With 128 bits, SFSR achieves the accuracy of 85.74%, which is higher than the second one (MSSRC) by almost 5%. Compared with the mostly recent deep learning method (i.e., DRM), the accuracy improvement of our method is more than 10%. Our results are also consistent with the results in face recognition (Schroff, Kalenichenko, and Philbin 2015). That is, one can make use of compact binary codes to achieve good recognition performance for face recognition problems although the main challenging problem in the YouTube Celebrities Dataset is incurred by tracking and detection errors. Compared with state-of-the-art image set classification methods with floating features, the proposed method can not only reduce the number of samples and features but also generate discriminative binary codes.

4 Conclusion

This paper has introduced a joint learning method for image-set classification that simultaneously learns compact binary codes and removes redundant samples. We seek a hashing function to generate binary codes containing large inter-class and small intra-class distances and then reduce redundant samples with discrete constraints in a low-rank way in two steps. Furthermore, to admit an analytical solution, the discrete objective function to a series of sub-problems is relaxed to obtain a high-quality discrete solution with a very low computational cost. A kernel method based on anchor points is further used to handle complex and nonlinear variations of images, which allows implicit and nonlinear feature mappings. At last, we formulate the binarization of an testing image set as a structured learning problem with discrete constraints by using the hashing function. The experimental results verified the efficiency and effectiveness of the proposed method in depressing and coding in large-scale face image sets. In the future, we will continue researching on hashing based sample reduction and representation method.

5 Acknowledgement

The authors would like to acknowledge the financial support including National Science Foundation of China (Grant No. 61473289, 61572463), the Youth Innovation Promotion Association CAS (Grant No. 2015190), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XD-

Methods	RNP(2013)	MSSRC(2013)	SGM(2015a)	PML(2015b)	DARG(2015)	DRM(2015)	SFSR
Accuracy	78.9%	80.75%	75.16%	70.32%	77.09%	72.55 %	85.74%

Table 1: Recognition rates of state-of-the-art image-set classification methods from the YouTube Dataset.

B02000000).

References

Chen, B.; Wang, J.; Zhao, H.; Zheng, N.; and Principe, J. C. Convergence of a fixed-point algorithm under maximum correntropy criterion. *IEEE Signal Processing Letters* 22(10):1723–1727.

Chen, Y.-C.; Patel, V. M.; Phillips, P. J.; and Chellappa, R. Dictionary-based face recognition from video. In *ECCV*.

Gong, Y., and Lazebnik, S. 2011. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*.

Grave, E.; Obozinski, G.; and Bach, F. 2011. Trace lasso: a trace norm regularization for correlated designs. In *NIPS*.

Gross, R., and Shi, J. 2001. The cmu motion of body (mobo) database. Technical report, Technical Report CMU-RI-TR-01-18, Robotics Inst.

Harandi, M. T.; Sanderson, C.; Shirazi, S.; and Lovell, B. C. 2010. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. In *CVPR*.

Hayat, M.; Bennamoun, M.; and An, S. 2015. Deep reconstruction models for image set classification. *IEEE TPAMI* 37(4):713–727.

He, R.; Cai, Y.; Tan, T.; and Davis, L. 2015. Learning predictable binary codes for face indexing. *Elsevier Pattern Recognition* 48(10):3160–3168.

He, R.; Tan, T.; and Wang, L. 2014. Recovery of corrupted low-rank matrix by implicit regularizers. *IEEE TPAMI* 36(4):770–783.

Hu, Y.; Mian, A.; and Owens, R. 2012. Face recognition using sparse approximated nearest points between image sets. *IEEE TPAMI* 34(10):1992–2004.

Huang, Z.; Wang, R.; Shan, S.; and Chen, X. 2015a. Face recognition on large-scale video in the wild with hybrid Euclidean-and-Riemannian metric learning. *Pattern Recognition* 48(10):3113–3124.

Huang, Z.; Wang, R.; Shan, S.; and Chen, X. 2015b. Projection metric learning on Grassmann manifold with application to video based face recognition. In *CVPR*.

Kim, T.; Arandjelovic, O.; and Cipolla, R. 2007. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE TPAMI* 29:1005–1018.

Kim, M.; Kumar, S.; Pavlovic, V.; and Rowley. 2008. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*.

Kittler, J.; Ghaderi, R.; Windeatt, T.; and Matas, J. 2001. Face verification using error correcting output codes. In *CVPR*, 755–760.

Kong, W., and Li, W.-J. 2012. Double-bit quantization for hashing. In *AAAI*.

Lee, K.; Ho, J.; Yang, M.; and Kriegman, D. 2003. Video-based face recognition using probabilistic appearance manifolds. In *CVPR*.

Liu, W.; Wang, J.; Ji, R.; Jiang, Y.; and Chang, S. 2012. Supervised hashing with kernels. In *CVPR*, 2074–2081.

Lu, J.; Wang, G.; Deng, W.; and Moulin, P. 2014. Simultaneous feature and dictionary learning for image set based face recognition. In *ECCV*.

Nishiyama, M.; Yuasa, M.; Shibata, T.; Wakasugi, T.; Kawahara, T.; and Yamaguchi, O. 2007. Recognizing faces of moving people by hierarchical image-set matching. In *CVPR*.

Ortiz, E. G.; Wright, A.; and Shah, M. 2013. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *CVPR*.

Rastegari, M.; Farhadi, A.; and Forsyth, D. 2012. Attribute discovery via predictable discriminative binary codes. In *ECCV*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*.

Shen, F.; Shen, C.; Liu, W.; and Shen, H. T. 2015. Supervised discrete hashing. In *CVPR*.

Wang, R.; Shan, S. G.; Chen, X. L.; and Gao, W. 2008. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*.

Wang, K.; He, R.; Wang, W.; Wang, L.; and Tan, T. 2013. Learning coupled feature spaces for cross-modal matching. In *ICCV*.

Wang, J.; Shen, H. T.; Song, J.; and Ji, J. 2014. Hashing for similarity search: A survey. *arXiv:1408.2927*.

Wang, W.; Wang, R.; Huang, Z.; Shan, S.; and Chen, X. 2015. Discriminant analysis on Riemannian manifold of gaussian distributions for face recognition with image sets. In *CVPR*.

Yang, M.; Zhu, P.; Gool, L. V.; and Zhang, L. 2013. Face recognition based on regularized nearest points between image sets. In *Automatic Face and Gesture Recognition*.

Zhang, G.; He, R.; and Davis, L. S. 2014. Simultaneous feature and dictionary learning for image set based face recognition. In *ACCV*.