Pattern Recognition 75 (2018) 4–14

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Learning structured ordinal measures for video based face recognition

Ran He [a,b,*], Tieniu Tan [a,b], Larry Davis [c], Zhenan Sun [a,b]

[a] The National Laboratory of Pattern Recognition and CAS Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), CASIA, Beijing 100190, China
[b] University of Chinese Academy of Sciences (UCAS), Beijing 100049, China
[c] The institute for advanced computer studies and the department of computer science, University of Maryland, College Park, MD 20742, United States

ARTICLE INFO

Article history:
Received 24 August 2016
Revised 13 December 2016
Accepted 2 February 2017
Available online 3 February 2017

MSC:
00-01
99-00

Keywords:
Ordinal measure
Metric learning
Local feature

ABSTRACT

Handcrafted ordinal measures (OM) have been widely used in many computer vision problems. This paper presents a structured OM (SOM) method in a data driven way. SOM simultaneously learns ordinal filters and structured ordinal features. It leads to a structural distance metric for video-based face recognition. The SOM problem is posed as a non-convex integer program problem that includes two parts. The first part learns stable ordinal filters to project video data into a large-margin ordinal space. The second seeks self-correcting and discrete codes by balancing the projected data and a rank-one ordinal matrix in a structured low-rank way. Weakly-supervised and supervised structures are considered for the ordinal matrix. In addition, as a complement to hierarchical structures, deep feature representations are integrated into our method to enhance coding stability. An alternating minimization method is employed to handle the discrete and low-rank constraints, yielding high-quality codes that capture prior structures well. Experimental results on three commonly used face video databases show that our SOM method with a simple voting classifier can achieve state-of-the-art recognition rates using fewer features and samples.

1. Introduction

Video-sharing websites are a fast-growing platform that allows internet users to distribute their video clips. There are often a large number of face videos in these websites. How to index, retrieve, and classify these face videos has become an active research topic in the area of video-based face recognition (VFR) [1]. Current VFR methods often perform recognition based on hundreds or thousands of floating point features [2], and store almost every face sample from a video clip. Since there can be (many) thousands of face samples in a video clip, high-dimensional dense features and large-scale registered samples result in tremendously large time and space complexity, which becomes a computational bottleneck when applying VFR methods to video-sharing websites.

Recently, binary code representations have drawn much attention in biometric recognition [3–5] and large scale image retrieval [6–8]. Among these binary coding methods, codes constructed from ordinal measures (OM) are one representative method. Ordinal measures [9] are common in human perceptual judgments. It is easy and natural for humans to rank or order the heights of two persons, although it is hard to estimate their precise differences [10]. Ordinal measures were originally used in social science [9] and then introduced to computer vision.

In biometrics, an OM is defined as the relative ordering of some property - for example, the average brightness of two adjacent regions (with 1 coding $A \succ B$ and 0 coding $A \prec B$) or the relative ordering of two color channels within the same region. Ordinal filters with a number of tunable parameters, are methods to analyze the ordinal measures of image features. The Haar wavelet and quadratic spline wavelet can be regarded as typical ordinal filters. Ordinal features are the binary codes of image features obtained by thresholding ordinal filters. Fig. 1 plots a simple illustration of OM.

In prior work, the set of handcrafted ordinal filters is chosen to correspond to some family of coherent patterns - like Gabor filters. The space of ordinal filters can therefore be quite large as the tunable parameters - scale, frequency, orientation - are varied, each giving rise to a potential ordinal feature. Different feature selection methods [10–12] have been used for OM to select a stable subset from the over-complete ordinal features. The term 'stable' indicates that the floating point features generated by an ordinal filter from the same class are expected to have large margins so that the corresponding ordinal features (binary codes) are robust to intra-class variations during binarization.

* Corresponding author.
E-mail address: rhe@nlpr.ia.ac.cn (R. He).

http://dx.doi.org/10.1016/j.patcog.2017.02.005
0031-3203/

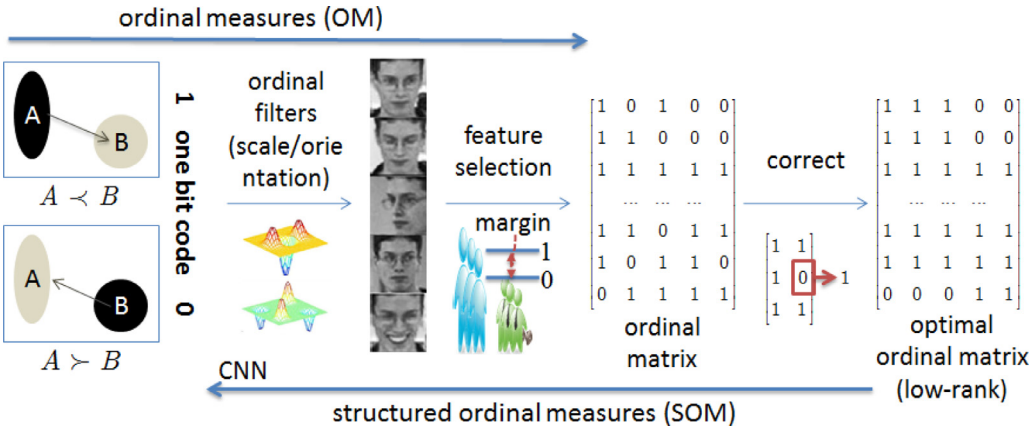


Fig. 1. An illustration of structured ordinal measures. Ordinal measure of visual relationship between two regions [10,11]. Previous OM methods apply feature selection methods to select over-complete ordinal features (binary codes) that are generated by handcrafted ordinal filters. SOM simultaneously seeks ordinal filters and optimal ordinal features in a data-driven way, makes the learned features low-rank and enforces an optimal ordinal matrix for classification. In SOM, one binary code of a sample can be corrected according to the codes of similar samples.

Motivated by the success of OM in iris [11], palmprint [10] and face recognition [3], we present what we refer to as a structured ordinal measure (SOM) method for video-to-video face recognition. Different from previous handcrafted OM methods, SOM simultaneously learns ordinal filters (SVM's) and structured ordinal features (binary codes) from video data as shown in Fig. 1. Considering that face appearances in video clips contain several facial variations and are similar in adjacent frames, we design the ordinal features of SOM to be stable and self-correcting binary codes. Stability indicates that the learned ordinal features are required to have large margins and to be clustered. The self-correcting character indicates that binary code of one frame depends not only on its corresponding ordinal filter (or coding function) but also on the binary values of similar (typically nearby in time) face samples. Because face images in a video clip often lie in a union of multiple linear subspaces [13,14], the features (binary code) assigned to the subset of faces from a single linear subspace should be similar. These binary codes can be potentially corrected by each other through a low-rank constraint on the matrix of constructed codes. One of the main advantages of our method is that it simultaneously reduces the number of dense features and eliminates redundant samples.[1]

We will formulate the SOM problem as a non-convex integer program problem that mainly includes two parts. The first part learns stable ordinal filters to project video data into a space in which the filtered data are separable with a maximum margin. This can be viewed as an instance of maximum margin clustering (MMC) [15]. The second finds self-correcting binary codes by balancing the projected real-value data and a rank-one ordinal matrix in a structured low-rank way. Weakly-supervised and supervised structures are considered for the ordinal matrix. We also integrate CNN feature representations into our method to enhance stability. An alternating optimization method provides an efficient discrete solution to deal with the discrete and low-rank constraints imposed on binary ordinal features. In addition, a simple voting classifier with a self-correcting process is proposed to efficiently compress and classify video clips. Experimental results on three commonly used face video databases show that our SOM method can achieve state-of-the-art recognition results using fewer features and samples. Compared to previous binary coding methods for still images (face or iris), SOM more efficiently utilizes the low-rank property of video data and hence is potentially useful for VFR problems.

There are three major contributions of this work:

(1) By employing the optimal ordinal matrices as output structures, SOM encourages ordinal features from the same class to have similar binary codes. To the best of our knowledge, SOM is the first algorithm that learns binary codes (or hashing) using output structures.
(2) Assuming that face images of a video clip are similar and related, we propose a self-correcting method to discretely binarize both gallery and probe videos. Our method utilizes the continuous information in videos and hence is effective for VFR tasks.
(3) As a by-product of SOM, we show that using a simple voting classifier improves over competing and complex classification models on fine grained datasets like the YouTube Celebrities dataset and offers an impressive compression ratio of CNN floating point features (20% face samples and 64-bit binary codes).

The rest of this paper is organized as follows. We briefly review some recent advances on binary coding methods in Section 2. In Section 3 and Section 4, we present the details of SOM and the optimal ordinal matrices respectively. Section 5 provides experimental results, prior to summary in Section 6.

2. Related work

Since OM methods are an instance of binary appearance features, we briefly review some recent advances on binary coding methods.

2.1. Biometric recognition

In biometrics, binary feature representation methods often focus on directly computing local image patches by the filters to generate binary codes. Local binary patterns (LBP) and ordinal measures are two representative binary features. There are many variations of these two features [3,4]. The definition and properties of OM in the context of biometrics can be found in [11].

Although OM's has been successfully applied to biometrics [16,17], there are still two open issues for OM. The first issue is the design of ordinal filters. The existing ordinal filters are often handcrafted. But handcrafted ordinal filters are too simple to represent complex human vision structures [18]. In addition, to improve stability and accuracy, these filters often contain a large number of

[1] Getting rid of redundant samples is important during both training and testing. In a video clip, the face can remain unchanging for long periods of time and that would bias the models towards that appearance.

parameters based on distance, scale and location, resulting in a potential feature set of OM. This naturally leads to the second issue, i.e., how to select the optimal set of ordinal features. Although various feature selection methods [10–12] have been employed to improve selection results, it is still difficult for a feature selection algorithm to select the optimal set from the over-complete set of OM.

Recently, data-driven binary feature methods, which learn local image filters from data, have drawn much attention. Cao et al. [19] utilized unsupervised methods (random-projection trees and PCA trees) to learn binary representations. Lei et al. [4] proposed a LBP-like discriminant face descriptor (DFD) by combining image filtering, pattern sampling and encoding. Chan et al. [20] combined cascade PCA, binary code learning and block-wise histograms to learn a deep network. Lu et al. [5] proposed a compact binary face descriptor (CBFD) to remove the redundancy information of face images. Although these methods indeed boost recognition performance on some challenging databases, their learned features are often high dimensional. For example, the dimensionality of histogram feature vectors of DFD and CBFD are 50,176 and 32,000 respectively. High dimensional and dense representations make these data-driven methods not applicable to VFR problems.

2.2. Image retrieval

Learning binary codes ('hashing') has been a key step to facilitate large-scale image retrieval. In image retrieval, the terminology 'hashing' refers to learning compact binary codes with Hamming distance computation. Similarity-sensitive hashing or locality-sensitive hashing algorithms [21,22], graph-based hashing [23], semi-supervised learning [24], support vector machine [25,26], Riemannian manifold [27], decision trees [7] and deep learning [28,29] have been studied to map high-dimensional data into a low-dimensional Hamming space. The authors in [23,26] argued that the degraded performance of hashing methods is due to the optimization procedures used to achieve discrete binary codes. Hence [23,26] tried to enforce binary constraints to directly obtain discrete codes [23,26]. A brief review of hashing methods for image search can be found in [28,30].

These hashing methods are often used for image search and retrieval but they may not achieve the highest accuracy for VFR problems. For example, the constraints in [23] maximize the information from each binary code over all the samples in a training set. However, adjacent face samples in a video clip often have nearly the same appearance so that these samples can have similar binary codes. In addition, to the best of our knowledge, there is no existing hashing methods that address image-set problems [31].

3. Structured ordinal measures (SOM)

3.1. Motivation

Consider a training set X from C classes, which consists of n biometric samples x_j $(1 \le j \le n)$ in a high dimensional Euclidean space R^d. The goal of previous OM methods is to identify ordinal filters over X to nonlinearly map each x_j to m ordinal features (an m-bit binary code). Since ordinal filters typically have a number of tunable parameters and so determine a huge set of possible ordinal features, various feature selection methods have been used to select the m ordinal features. The selected ordinal features of all samples form a binary matrix $B = [b_1, \ldots, b_n] \in R^{m \times n}$, referred to as an **ordinal matrix**. Previous OM methods select ordinal filters one by one (using a greedy approach) and hence neglect the output structure of ordinal features. For example, video data are often low-rank.

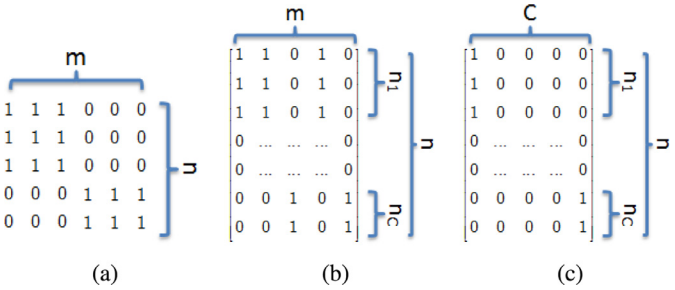


Fig. 2. Three types of the optimal ordinal matrices. (a) The optimal ordinal matrix for a two-class problem. (b) Weakly-supervised ordinal matrix constructed via appearance information. Binary codes of all samples from the same class are arbitrary but unique and identical. (c) Supervised ordinal matrix via the spectral matrix of linear discriminant analysis [36].

In biometrics, since intra-class variations of biometric samples are often very large [32], good ordinal measures should generate similar binary codes for the samples from one subject. In addition, a large difference between two quantities will result in more stable binary features. For example, the greater the color difference between two image regions, the more easily humans order their relative brightness (1 or 0); and the greater the height difference between two persons, the more easily humans rank their relative heights.

To obtain stable ordinal features, we introduce the following minimization problem for OM,

$$\min_{W,\xi,B} \mu\xi + \lambda_1 \|W\|_2 + \sum_c \|B^c\|_* \qquad (1)$$
$$s.t.\ B_{ij}(w_i^T X_j) \ge 1 - \xi_{ij},$$
$$\xi_{ij} \ge 0,\ B_{ij} \in \{-1, 1\}$$

where μ and λ_1 are constants, and $\|.\|_*$ denotes the matrix trace norm (i.e., the sum of its singular values). B^c represents all ordinal features from the cth class [33]. The parameter matrix $W = [w_1, \ldots, w_m] \in R^{d \times m}$ represents a set of ordinal filters. As defined in Section 2, a parameter matrix W contains a set of ordinal filters only if W can result in consistent orders for the samples from the same class, e.g., $W^T X$ generates an ordinal matrix as in Fig 2. In contrast to the binary coding methods [4,5,11] that are based on local image patches, (1) directly uses the whole image as an input to find compact codes[2]. More important, (1) aims to simultaneously seek ordinal filters (W) and optimal ordinal features (B).

The low-rank constraint in (1) encourages the ordinal features from the same class to be correlated. This constraint reduces the redundancy of video data and corrects some binary codes whose corresponding values ($W^T X$) are close to SVM's separating hyperplanes. We also want to enforce that the learned B is close to the optimal ordinal (binary) matrix for classification, resulting in the following minimization problem,

$$\min_{W,\xi,B} \mu\xi + \lambda_1 \|W\|_2 + \sum_c \|B^c\|_* + \lambda_2 \|B - S\|_F^2 \qquad (2)$$
$$s.t.\ B_{ij}(w_i^T X_j) \ge 1 - \xi_{ij}, \xi_{ij} \ge 0,\ B_{ij} \in \{-1, 1\}$$

where $S \in R^{m \times n}$ is a prior ordinal matrix that defines a desired output structure for ordinal features. We postpone discussion of the design of S until Section 4. Since the OM problem in (2) imposes an output structure on ordinal filter learning, we refer to the problem in (2) as learning a **structured ordinal measure**.

Even without the structured low-rank constraint, (2) is difficult to solve [15]. Unlike supervised SVM that can be formulated

[2] In face recognition, dividing a face image into small patches can capture nonlinear facial variations well and so improves recognition rates. The learned filters in (1) can also be applied to local patches as in previous binary coding methods.

as a convex optimization problem, (2), even without the structured low-rank constraint, is still a non-convex integer optimization problem. It is an instance of maximum margin clustering [15]. To simplify the minimization of (2), we relax (2) by introducing an equality constraint on B as follows,

$$\begin{aligned} \min \mu\xi + \lambda_1\|W\|_2 + \sum_c \|B^c\|_* + \lambda_2\|B-S\|_F^2 + \|E\|_F^2 \\ s.t.\ B = W^TX + E,\ B_{ij} \in \{-1,1\}, \\ B_{ij}(w_i^TX_j) \geq 1-\xi_{ij}, \xi_{ij} \geq 0 \end{aligned} \tag{3}$$

where $E \in R^{m \times n}$ is an error term to reduce the loss during binarization. Since $\|B-S\|_F^2 = \sum_c \|B^c - S^c\|_F^2$, (3) actually seeks discrete binary codes by balancing floating point data W^TX and a rank-one ordinal matrix S^c in a structured low-rank way.

Our SOM formulation in (3) has two major advantages: 1) the introduction of the low-rank constraint and error term makes SOM more flexible during binarization. The learned binary codes depend on their corresponding floating point values as well as prior structures. Different from the binary codes that are directly generated by ordinal filters or hashing functions, the binary codes of SOM can be self-corrected by the structure constraints, resulting in self-correcting codes. 2) Since S^c is a rank-one matrix, λ_2 plays the role of controlling the number of learning samples. The rank-one matrix indicates that there is only one unique sample in this matrix. The larger the value of λ_2, the more B^c resembles S^c. In practice, the rank of B^c will be larger than one because a face video clip often contains several face variations.

3.2. Optimization

The optimization problem in (3) is a hard computational problem (non-convex integer optimization), which belongs to the class of maximum margin clustering problems [15]. Fortunately, we do not need to find the global minimum because local minima produce good ordinal features. Hence we can decompose the non-convex problem in (3) into subproblems as in MMC. A local minimum can be obtained by solving a series of SVM training and binary code learning problems. An overview of our iterative algorithm is as follows.

First, fixing variables B and E, we minimize (3) w.r.t. variables W and ξ, resulting in a multiple linear SVM problem in (4) (one for each ordinal feature) [34]. To learn the ith SVM,[3] the columns of X and the elements of the ith row of B are used as training data and labels respectively.

$$\begin{aligned} \min_{W,\xi} \mu\xi + \lambda_1\|W\|_2 \\ s.t.\ B_{ij}(w_i^TX_j) \geq 1-\xi_{ij}, \xi_{ij} \geq 0 \end{aligned} \tag{4}$$

Second, fixing variables W and ξ, (3) takes the following form w.r.t. B and E,

$$\begin{aligned} \min_{B,E} \sum_c \|B^c\|_* + \lambda_2\|B-S\|_F^2 + \|E\|_F^2 \\ s.t.\ B = A + E,\ B_{ij} \in \{-1,1\} \end{aligned} \tag{5}$$

where $A = W^{\{t+1\}T}X$. By substituting the equality constraint into the objective function of (5), we can reformulate (5) as follows,

$$\begin{aligned} \min_B \|A-B\|_F^2 + \sum_c \|B^c\|_* + \lambda_2\|B-S\|_F^2 \\ s.t.\ B_{ij} \in \{-1,1\} \end{aligned} \tag{6}$$

Since $||.||_F^2$ is separable, the solution of (6) can be independently obtained by minimizing the following subproblem for each class c,

$$\begin{aligned} \min_{B^c} \|A^c - B^c\|_F^2 + \|B^c\|_* + \lambda_2\|B^c - S^c\|_F^2 \\ s.t.\ B_{ij}^c \in \{-1,1\} \end{aligned} \tag{7}$$

[3] The ℓ_1 regularized linear SVM is implemented by LIBLINEAR: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

To minimize the low-rank problem in (7), we first need to introduce a variational formulation for the trace norm [35],

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

$$\|B\|_* = \frac{1}{2}\inf_{L \geq 0} tr\left(B^TL^{-1}B\right) + tr(L) \tag{8}$$

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

Using this lemma, we can reformulate (7) as,

$$\begin{aligned} \min_{B^c} \min_{L \geq 0}\ \|A^c - B^c\|_F^2 + tr(B^{cT}L^{-1}B^c) \\ +\lambda_2\|B^c - S^c\|_F^2 + tr(L)\ s.t.\ B_{ij}^c \in \{-1,1\} \end{aligned} \tag{9}$$

The problem in (9) can be alternately minimized. When L is fixed, we can use the discrete cyclic coordinate descent method to obtain B^c bit by bit. For simplicity, we develop a simple and direct method to find B^c. That is, disregarding the integer constraint, the solution of B^c takes the following form by setting the derivative of (9) w.r.t. B^c equal to zero,

$$B^c = ((1+\lambda_2)I + L^{-1}) \backslash (A^c + \lambda_2 S^c). \tag{10}$$

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 VFR. Algorithm 1 summarizes the procedure to learn structured ordinal filters. λ_2 is set to 0.1 throughout this paper.

Algorithm 1: Learning structured ordinal filters.

Input: Data matrix $X \in R^{d \times n}$ and ordinal matrix $S \in R^{m \times n}$
Output: Ordinal Filters $W \in R^{d \times m}$

```
repeat
    Train m linear-SVMs to update W using B^{t-1} as training labels.
    Compute A = {W^t}^T X.
    repeat
        Compute L = (B^c B^{cT})^{1/2}.
        Compute B^c via (10).
        Let B^c = sgn(B^c).
    until The variation of B is smaller than a threshold.
    t=t+1.
until The variation of B is smaller than a threshold.
```

3.3. Classification

When applying SOM (or binary code learning methods) to biometric recognition, SOM must generate ordinal features for any data sample beyond the sample points in the training set X. Given a new probe dataset X^p, a hashing algorithm H with parameter W typically applies the sign function $sgn(.)$ to the hashing function $f_W^H(X^p)$ to obtain the binary codes [23,26], i.e., $B^H = \text{sgn}(f_W^H(X^p))$.

VFR can be viewed as an image-set classification/retrival problem [31]. The samples in a probe (or gallery) dataset are from a video clip and so have a low-rank structure. Hence, instead of using the sign function, we propose a low-rank method to construct the binary codes for a probe video as follows,

$$\begin{aligned} \min_B \left\{\|E\|_F^2 + \|B\|_*\right\} \\ s.t.\ B = f_W^H(X^p) + E,\ B_{ij} \in \{-1,1\} \end{aligned} \tag{11}$$

Compared to directly using the sign function $sgn(.)$ to obtain binary codes, (11) utilizes a low-rank prior to find binary codes. This makes the binary codes B not only depend on the function

$f_W^H(.)$. The values in B can be potentially changed (or corrected) by each other due to the low-rank constraint. (11) is a sub-problem of (7) when λ_2 is set to zero. Hence (11) can be alternatively minimized as (7).

Given the binary codes constructed from (11), a simple nearest neighbor classifier for each unique code in B (since many samples can be mapped to the same code by the optimization) with voting is used as classifier to report recognition rates. The class label of the majority class in a video sequence is taken as the final class label of this sequence. In addition, since the low-rank constraint in (11) tends to make the column samples in B correlated, it also tends to reduce the number of different samples in B. We introduce the term **compression ratio of samples** for VFR, i.e., compression ratio = the number of unique samples/the total number of samples. A lower compression ratio of an algorithm indicates that the algorithm needs less storage space (and as a consequence less computational time).

In addition, since there is no a rank-one constraint in (11) (compared to (2)), compression ratio will tend to be high as the number of desired bits increases. If some priors of the rank of a video clip are given or a lower compression ratio is required, we can further impose a rank constraint on (11), resulting in the following minimization problem,

$$\min_B \left\| f_W^H(X^p) - B \right\|_F^2 \quad (12)$$
$$s.t.\ rank(B) \le r,\ B_{ij} \in \{-1, 1\}$$

where $rank(.)$ is the matrix rank operator and r is constant. The rank constraint in (12) makes the rank of B is smaller than r. That is, all binary samples can be linearly represented by r binary vectors. As a result, the number of unique samples is potentially related to r.

4. Ordinal matrices for classification

In this section, we discuss the design of the optimal ordinal matrices in (2). Then we discuss combining deep feature representation to improve the stability of SOM.

4.1. The optimal ordinal matrix

We begin the study of the optimal ordinal matrix S for (2) with a two-class problem. We expect that all intra-class and inter-class sample pairs of binary codes are well separated with a large margin, i.e.,

$$J(B) = \frac{1}{\mu_1} \sum_{c_i \neq c_j} \left\| b_i - b_j \right\|_0 - \frac{1}{\mu_2} \sum_{c_i = c_j} \left\| b_i - b_j \right\|_0 \quad (13)$$

where $B = [b_1, \ldots, b_n] \in R^{m \times n}$ is a binary matrix, μ_1 and μ_2 are the numbers of extra-class and intra-class pairs respectively, and $\|.\|_0$ is the counting norm (i.e., the number of nonzero entries in a vector or matrix). Each row of B^T corresponds to the binary code of one data item. The first term of (13) rewards items from difference classes having large Hamming distance, while the second term penalizes items from the same class having small Hamming distance. The maximization of $J(B)$ is NP-hard. By analyzing $J(B)$, we make the following two observations on its optimal solution,

Proposition 1. *The maximum value of $J(B)$ is equal to the number of bits (m), i.e.,* $\max_B J(B) \le m$.

Proof. According to the definition of the ℓ_0 norm, we can easily derive that $\max_B(J(B)) < m$. In addition, when $\hat{B}$ satisfies,

(a) For $\forall i, j, k$ and $c_i \neq c_j$, if $b_{ik} \neq b_{jk}$, then $\frac{1}{\lambda_1} \sum_{c_i \neq c_j} \left\| b_i - b_j \right\|_0 = m$;

(b) For $\forall i, j, k$ and $c_i = c_j$, if $b_{ik} = b_{jk}$, then $\frac{1}{\lambda_2} \sum_{c_i = c_j} \left\| b_i - b_j \right\|_0 = 0$,

we obtain $J(\hat{B}) = m$ (Fig. 2(a) gives an example of $\hat{B}$). Hence $\max_B J(B) \le m$. □

Proposition 2. *If there exists a $\hat{B}$ such that $J(\hat{B}) = m$, the $\hat{B}$ satisfies the following two conditions. (a) All the samples in each class have a unique binary code. (b) The sample code of one class is orthogonal to that of the other class.*

Proof. If $c_i = c_j$ and $b_{ik} \neq b_{jk}$, then $\sum_{c_i = c_j} \left\| b_i - b_j \right\|_0 > 0$ so that $J(B) < m$. Since $b_{ik} \in \{0, 1\}$ and $\left\| b_i - b_j \right\|_0 = m$ for $c_i \neq c_j$, $b_i^T b_j = 0$. Hence b_i is orthogonal to b_j when $c_i \neq c_j$ and $J(B) = m$.

From Propositions 1 and 2, we can easily obtain the optimal ordinal matrix for a two-class problem as shown in Fig. 2(a). Previous ordinal feature selection methods [10,11] actually select ordinal filters one by one so that the selected filters generate codes like in Fig. 2 (a). When there are multiple classes, the problem of determining the optimal binary codes becomes complex. Inspired by Propositions 1 and 2, we consider two types of ordinal matrices to approximate the optimal ordinal matrix (shown in Fig. 2(b) and (c)).

For the weakly-supervised ordinal matrix, we just require that the binary codes of each class be unique. There are many ways to generate informative binary codes for this case, e.g., random binary codes and Hadamard codes [37]. Since ordinal filters perform learning based on human face appearances, we also expect that the weakly-supervised ordinal matrix would capture useful appearance information of video data. To accomplish this, we apply the unsupervised version of Iterative Quantization (PCA-ITQ) [38] to the mean faces of each class to generate the corresponding unique binary code for each class. Then, the weakly-supervised ordinal matrix contains appearance information while the binary codes of different classes are largely uncorrelated.

For the supervised ordinal matrix, we simply employ the spectral matrix of linear discriminant analysis [36] (the regression target of multi-class linear regression). In this spectral matrix, the binary codes of the samples from any one class have just one bit set, which define the orders of a class. Since this spectral matrix contains discriminative information, the ordinal matrix will contain supervised information if this spectral matrix is used as the ordinal matrix. However, the code length of this spectral matrix can be only C. If code lengths larger than C are needed, we can obtain longer binary codes by combining the spectral matrix with the weakly-supervised ordinal matrix.

4.2. Deep feature representations

Since there are large variations of intra-class samples in uncontrolled VFR environments, it is often difficult to use one type of local appearance features to obtain satisfactory recognition results. Hence, biometric researchers often combine several local feature to improve generalization ability and recognition performance. In [39], Gabor and LBP were combined to enhance the representation power of the spatial histogram. In [3], Gabor ordinal measures were proposed to improve distinctiveness of Gabor features and robustness of OM's. In [4,20,40,41], different techniques are combined together to achieve state-of-the-art results.

Inspired by the success of the combination of several appearance features, we couple SOM with deeply learned features from convolutional neural networks (CNN) [42] to improve coding stability. Benefiting from CNN's deep architecture and supervised learning approach [43–45], CNN's can efficiently deal with large amounts of data and generate a hierarchical and discriminative feature representation. The use of deeply learned features makes the

learned ordinal features contain not only the prior structure from data but also the hierarchical structure of local image patches.

The CNN network implemented by Alex ('cuda-convnet')[4] is used as our deep architecture. This CNN first feeds gray scale images to two convolutional layers, each followed by a normalization layer and a max-pooling layer. Then, two locally connected layers are connected to the output of the second max-pooling layer, and finally to a C-way soft-max regression layer (C is the number of classes) that produces a distribution over class labels. The input to this network is each cropped gray scale face image in a video without any preprocessing. The last C-way soft-max regression layer provides supervised information for learning face representations. The outputs of the last locally connected layers (fc6) before the softmax loss are employed as deep feature representations.

The testing time of the proposed methods consists of two parts. The first part is the computational time of CNN feature extraction. We can resort to a light CNN model [44] to reduce computation time. The second part involves the computation to acquire binary codes. If we only apply the linear projection matrix W in Eq. (2) to obtain binary codes, the computation complexity is $O(d \times m \times n)$. Moreover, if Eq. (12) is used to obtain binary codes, we can resort to the fast algorithm [46] to find a low-rank approximation.

5. Experiments

In video-sharing websites, there are a large number of face videos, each of which contains hundreds of face images. Using binary features to represent these face images will significantly save computational power and storage space. Hence, VFR is a good test platform to evaluate SOM. All experiments are run 10 times by repeating the random selection of training/testing set. For all binary code methods, the simple nearest neighbor classifier for each unique code in the probe set with voting is used as a classifier to report recognition rates.

5.1. Methods

We systematically compare SOM with popular techniques from three categories. **SOM1** and **SOM2** indicate Algorithm 1 using the last two structures from Fig. 2(b) and (c) respectively. For SOM2, the bits from the optimal matrix for SOM1 is appended to that for SOM2 as discussed in Section 4 if code length is larger than the number of classes.

For the first category, we compare SOM with state-of-the-art data-driven binary feature methods in biometrics, including discriminant face descriptor (DFD) [4], Gabor ordinal measures (GOM) [3], and compact binary face descriptor (CBFD) [5]. As in [5], cosine distance is used for the three methods to achieve their best recognition accuracy. Since the feature dimensions of DFD and CBFD are too high, whitened PCA (WPCA) is applied to reduce their feature dimensions to 1000 [5].

For the second category, we compare SOM with popular hashing methods, including locality sensitive hashing (LSH) [47], iterative quantization (ITQ) [38], kernel-based supervised hashing (KSH) [6], fast supervised hashing (FastH) [7], and supervised discrete hashing (SDH) [26]. For ITQ, its supervised version (CCA-ITQ) and unsupervised version (PCA-ITQ) are included. PCA is used as a preprocessing step for CCA-ITQ. For SDH, we use the notation SDH-n to indicate that SDH uses image pixels rather than nonlinear RBF kernel mapping as its input. Hamming distance is computed on each pair of face samples in training/testing sets.

For the last category, we compare SOM with popular VFR methods, including discriminative canonical correlations (DCC) [48], manifold discriminant analysis (MDA) [49], sparse approximated nearest point (SANP) [50], sparse representation for video (SRV) and its kernelized version KSRV [13], covariance discriminative learning (Cov+PLS) [51], jointly learning dictionary and subspace structure (JLDSS) [14], image sets alignment (ImgSets) [31], regularized nearest points (RNP) [52], and mean sequence sparse representation-based classification (MSSRC) [53]. As in [13,14,52,53] , we directly cited the best recognition rates of these methods from the literature.

5.2. Databases

Three commonly used face video datasets are used to evaluate different methods, including,

The Honda/UCSD dataset [54] is composed of 59 video sequences of 20 subjects. The sequences of each subject contain pose and expression variations. The lengths of the sequences vary from 12 to 645. Fig. 3(a) shows cropped images from this dataset. We follow the standard training/testing configuration in [14,49–51]: 20 sequences are used for training and the remaining 39 sequences for testing. All video frames are used to report classification results. Since there are only 39 testing sequences, the improvement of recognition rates is 2.6% ({1/39}*100%) when one additional sequence is correctly classified.

The Mobo (Motion of Body) dataset [55] was originally published for human pose identification. It contains 96 sequences of 24 different subjects walking on a treadmill. Each subject has four video sequences corresponding to four walking patterns respectively. These patterns (slow, fast, inclined, and carrying a ball) were captured using multiple cameras. Fig. 3(b) shows some cropped images from three subjects. We follow the standard training/testing configuration in [14,49–51]. One video was randomly chosen as training and the remaining three for testing. The improvement of recognition rates is (1.4% = 1/72*100%) if one additional video sequence is correctly classified.

The YouTube Celebrities dataset [56] contains 1910 video clips of 47 human subjects (actors, actresses, and politicians) from the YouTube website. Roughly 41 clips were segmented from 3 unique videos for each person. These clips are mostly low resolution and highly compressed. Each facial image is cropped to size 30 × 30 as shown in Fig. 3(c). This dataset is challenging because it contains large facial variations (e.g., pose, illumination and expressions) and tracking errors in the cropped faces. Following the standard setup, 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.

5.3. Algorithmic analysis

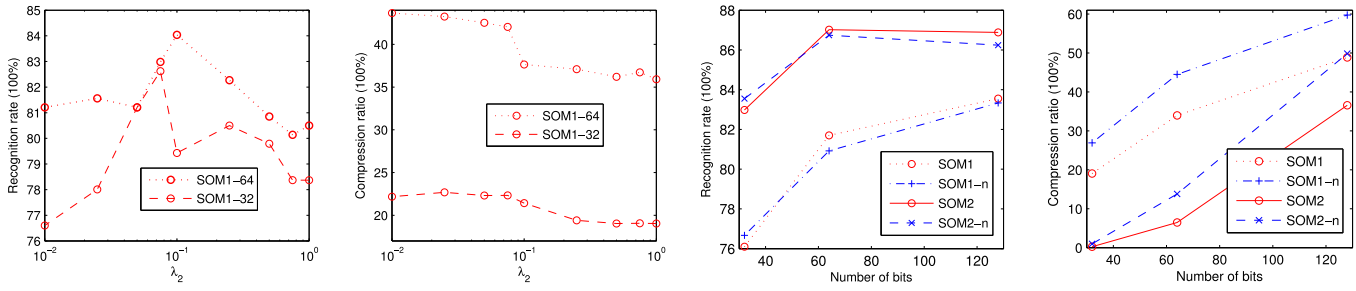
Since our SOM method consists of several parts to improve performance, we investigate the effectiveness of each part on the YouTube Celebrities dataset. To simplify parameter setting, we directly use the default parameter setting of μ and λ_1 in the LIBLINEAR SVM source code. Hence there is only one parameter λ_2 to control the effectiveness of output structures.

Fig. 4(a) and (b) show recognition rates and compression ratios of samples as a function of λ_2 respectively. Experimental results are from one single run. The lower compression ratio of an algorithm is, the better the algorithm is. We observe that parameter λ_2 affects both recognition rates and compression ratios. When λ_2 is a large, the output structure term $\|B - S\|_F^2$ dominates (5). If λ_2 is sufficiently large, the optimal solution of B will equal the ordinal matrix S, which indicates directly using S as the class labels of SVM to perform binary code learning. When λ_2 tends to be zero,

[4] https://code.google.com/p/cuda-convnet/.

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

Fig. 3. Cropped facial images of three different subjects in the three video databases respectively.



(a) (b) (c) (d)

Fig. 4. Recognition rates and compression ratios of SOM under different parameter setting. (a) Recognition rates as a function of λ_2. (b) Compression ratios of samples as a function of λ_2. (c) Average recognition rates with or without (11). SOM-n indicates that the SOM method without using (11). (d) Average compression ratios of samples with or without (11).

(5) becomes maximum margin clustering [15]. That is, we seek a global ordinal filter matrix W to group the samples from the same class into several clusters.

Since S^c is a rank-one matrix, B will be a rank-one matrix if B is equal to S. In VFR problems, a video clip often contains many face variations so that it is difficult to use one binary vector to represent all face variations. From Fig. 4(b), we also observe that the rank of the learned B is larger than 1. Hence, to keep the diversity of learned B, it is not a good strategy to directly use S as the class labels of SVM or to set λ_2 to a large value, although a larger λ_2 will result in better compression. Meanwhile, setting λ_2 too small will also damage performance. If λ_2 tends to zero, there will be no structure constraints to ensure that the learned ordinal features are similar to the optimal ordinal matrix for classification. Hence, the performance of SOM will decrease in terms of both recognition rates and compression ratios.

Fig. 4(b) and (c) show recognition rates and compression ratios of samples without using (11) respectively. SOM-n indicates that the SOM method uses *sgn*(.) function to obtain binary codes rather than using (11). We observe that using (11) further improves recognition rates and reduces compression ratios. This indicates that our SOM methods can correct some binary codes such that the learned codes become correlated. Since video data often contain a large number of face samples, it is impossible to make face samples uncorrelated as assumed by hashing methods. Reducing the redundancy of video data should be helpful for performance. We also observe that the improvement using (11) is not significant. We regard these results as reasonable because CNN features have powerful ability to learn discriminative representations. Since the binary codes learned by SOMs are discriminative enough on CNN features, there is a limited potential to further improve performance.

5.4. Comparisons to binary code methods

Table 1 and Figs. 5–7 show recognition rates and compression ratios of different binary code learning methods on the three video face databases. From these results, we make several observations:

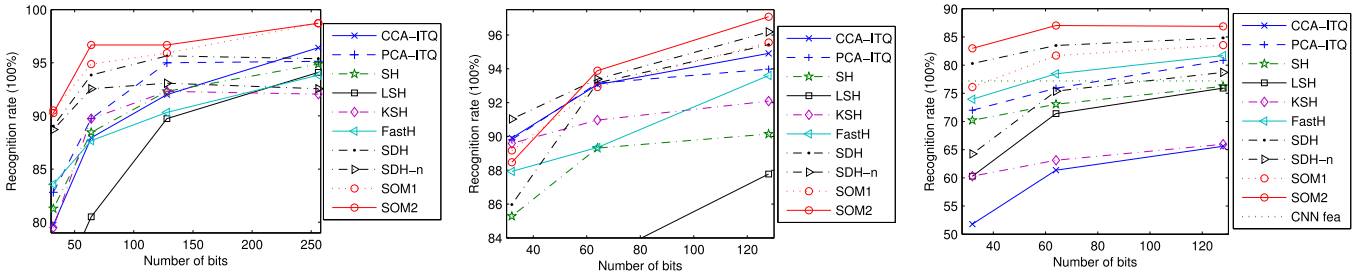
High-dimensional and dense features are powerful for VFR. Three binary feature representation methods (GOM, CBFD and DFD) obtain the highest recognition rate (close to 100%) on the Honda dataset, and comparable recognition rates on the other two datasets. However, the best recognition rates of these three methods are obtained by cosine distance rather than Hamming distance. Dense feature representations will result in very high computational costs for VFR. For the Honda dataset, we can see that longer codes will lead to better recognition rates. The recognition rates of CCA-ITQ, LSH, FastH, SOM1 and SOM2 increase quickly as the number of bits increases.

Compared to the hashing methods designed for image retrieval, SOM methods are more effective for VFR. On all three databases, SOM methods achieve the highest recognition rates, and consistently outperform their hashing competitors. This may be because SOM methods can utilize and preserve the structure information from face videos. Since SOM2 considers discriminative binary codes in its prior structure, SOM2 performs better than SOM1 on the last two databases. On the YouTube database, since CNN features capture face variations well, SOM methods obtain state-of-the-art recognition rates compared to the complex classification models (e.g., image set models). It should be noted that the results for these other models are not based on CNN features, and their performance should improve if they were applied to those features. More important, SOM methods use 64-bit binary features to obtain a better result than directly using CNN features in a nearest neighbor recognition framework, which offers an impressive compression ratio of 1152-dim CNN features.

Binary code learning methods provide a potential way to reduce the number of registered samples. Since there are many face samples in a video clip, a lower compression ratio of an algorithm indicates that the algorithm needs smaller storage space and computational time. Since PCA-ITQ and CCA-ITQ aim to quantize the face samples so that they are uncorrelated, they should learn different binary codes for different samples. However, their compression ratios on the training and testing sets are smaller than 100%. This indicates that there are some samples to have the same binary code, which makes the uncorrelated constraints work not well. In addition, compression ratios of different methods on the training set seem to be lower than those on the testing set. This indicates that there are large difference between the videos in the training and testing set so that the learned coding functions more accu-

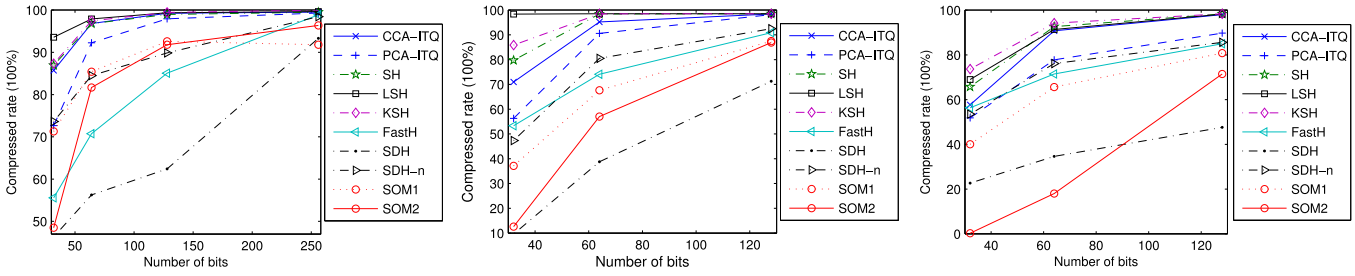
Table 1
Experimental results of three state-of-the-art binary feature representation methods. 'RR', 'CS1' and 'CS2' indicate recognition rate, compression ratio on the testing set, and compression ratio on the training set respectively.

Methods(dim)	Honda			Mobo			Youtube		
	RR	CS1	CS2	RR	CS1	CS2	RR	CS1	CS2
GOM(2560)	99.0%	100.0%	100.0%	92.6%	99.7%	100.0%	68.1%	99.3%	99.3%
CBFD(32000)	99.5%	99.4%	100.0%	95.1%	100.0%	100.0%	66.3%	99.3%	99.3%
DFD(50176)	99.2%	100.0%	100.0%	93.6%	100.0%	100.0%	64.7%	99.3%	99.3%



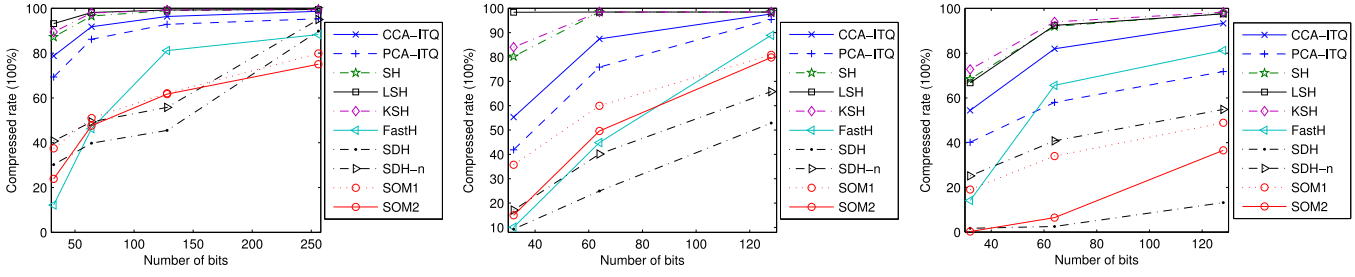
(a) Honda (b) Mobo (c) Youtube

Fig. 5. Recognition rates of different binary code learning methods.



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

Fig. 6. Compression ratios of different binary code learning methods on the three testing sets. Compression ratio = the number of unique samples/ the total number of samples. The lower compression ratio an algorithm has, the better the algorithm is.



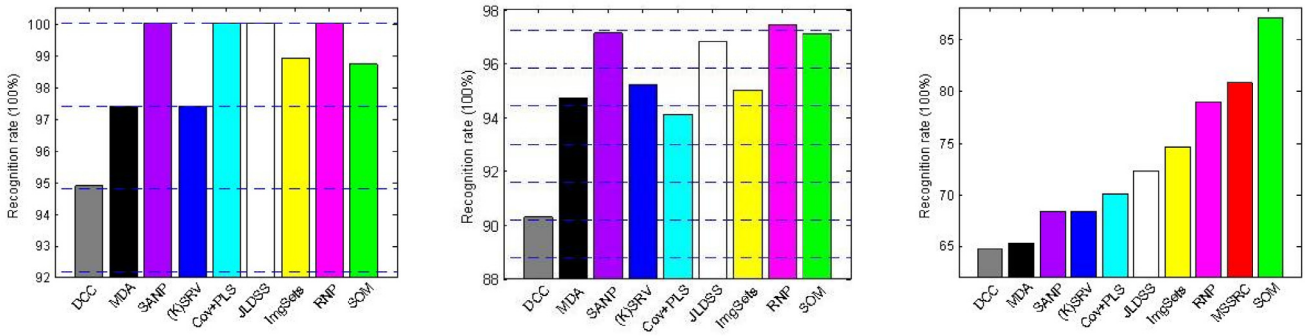
(a) Honda (b) Mobo (c) Youtube

Fig. 7. Compression ratios of different binary code learning methods on the three training sets.

rately capture the facial variations in the training set than those in the testing set.

FastH, SDH, SOM1 and SOM2 obtain lower compression ratios than other methods, which indicates that these methods can reduce intra-class variations. On the Honda and Youtube databases, SDH's performance seems to mainly benefit from its nonlinear RBF kernel mapping and anchor points, which forces the data to be similar to anchor points, resulting in low compression ratios. Without the nonlinear mapping, SDHn performs no better than other methods. Since the nonlinear RBF kernel mapping is an independent step for SDH, this data mapping can also be integrated into other methods as a preprocessing step if applicable. In contrast to SDH, SOM methods employ low-rank constraints to naturally group data to different clusters (or anchor points).

The optimal ordinal matrix for classification plays an important role for SOM. Although SOM1 and SOM2 are both minimized by Algorithm 1, they perform differently in terms of recognition rate and compression ratio. This is because SOM makes use of ordinal matrices as output structures that are helpful for classification. Different output structures result in different characteristic SOM's. Finding or defining the optimal ordinal matrix is still an open problem for ordinal measure and hashing. The coding theory from information theory [37] may provide useful insights for binary code learning methods.



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

Fig. 8. Recognition rates of different VFR methods on the three video databases. The interval between two dashed lines indicates the improvement of recognition rates if one additional video sequence is correctly classified.

5.5. Comparisons to VFR methods

In this subsection, we compare the proposed SOM methods with prevalent VFR methods that are based on hundreds of floating point features. Fig. 8(a) plots the average recognition rates of different VFR methods on the Honda dataset. The interval between two dashed lines indicates the improvement in recognition rates (2.6%) if one additional video sequence is correctly classified. The highest recognition rate achieved by SOM is 98.7% at 256 bits. We observe that the recognition rates of most of the compared methods are between 97.4% and 100%. This indicates that there is at most one misclassified sequence in the randomly selected subsets. These results also show that we can use only binary features and achieve state-of-the-art results on the Honda dataset.

Fig. 8(b) plots the average recognition rates of different VFR methods on the CUM Mobo dataset. The interval between two dash lines indicates the improvement of recognition rates (1.4% = 1/72*100%) if one additional video sequence is correctly classified. RNP achieves the highest recognition rate 97.4% ± 1.5%. In contrast, the recognition rate of SOM is 97.1%. This indicates that RNP outperforms SOM in some random selection cases but not in other cases. The reason is probably that SOM simply uses a nearest neighbor classifier with voting. Since SOM is a binary feature representation method and RNP is an image set method, we consider the result of SOM to be comparable to that of state-of-the-art VFR methods. In addition, an image set algorithm can also be applied to ordinal features to further improve accuracy.

Fig. 8(c) plots the average recognition rates of different VFR methods on the Youtube dataset. We observe that MSSRC and SOM are the two best methods on this data set. Their average recognition rates are 80.8% and 87.0% respectively. The accuracy improvement of SOM against MSSRC is more than 6%. The high accuracy of MSSRC is due to its robust tracker that successfully tracked 92% of the videos as compared to the 80% tracked by other methods. Since the low quality of video frames incurred by the high compression rate generates large tracking errors and noise in the cropped faces [50], a good tracker should significantly improve recognition accuracy. However, SOM did not use any preprocessing techniques (such as histogram equalization or an enhanced tracker). These results show that using a simple voting classifier can improve over the complex VFR models on the fine grained YouTube dataset. In addition, SOM can use a 64-bit representation to achieve a better recognition result than 1152-D floating point CNN representation, which offers an impressive compression ratio over CNN features.

6. Conclusion

We introduced the problem of designing data-driven ordinal structures for ordinal measures learning, and developed a structured ordinal measure method for video-based face recognition. By reformulating the problem in terms of an implied equivalence relation, we posed the learning problem as a non-convex integer program problem that mainly includes two parts. The first part learns stable ordinal filters to project video data into a large-margin ordinal space. The second seeks self-correcting and discrete codes by balancing the projected data and a rank-one ordinal matrix in a structured low-rank way. Weakly-supervised and supervised structures are considered for the ordinal matrix. We developed an alternating minimization method to efficiently minimize the proposed non-convex formulation. Experimental results demonstrate that our SOM methods provide state-of-the-art results with fewer features and samples on three commonly used video face databases.

The future work lies in two directions. First, our results show that the proposed output structures (the optimal ordinal matrices) are useful for video-based face recognition. Hence one direction is to design or learn optimal ordinal matrix based on various facial attributes, which have been shown to further improve recognition rates. Second, our results also show that SOM can efficiently compress redundant samples, resulting in a small set of unique samples. During classification, these unique samples can be treated as representative samples or anchor points to represent all video samples. Hence another potential direction is to apply the proposed method to the area of representative sample learning.

Acknowledgment

The authors would like to greatly thank Dr. Shu Zhang for implementing the codes of deep learning. This work is funded by State Key Development Program under (Grant No. 2016YFB1001001), and National Natural Science Foundation of China (Grant No. 61622310, 61473289).

References

[1] Z. Huang, R. Wang, S. Shan, X. Chen, Face recognition on large-scale video in the wild with hybrid euclidean-and-riemannian metric learning, Pattern Recognit. 48 (10) (2015) 3113–3124.
[2] Y. Chen, J. Su, Sparse embedded dictionary learning on face recognition, Pattern Recognit. 64 (2017) 51–59.
[3] Z. Chai, Z. Sun, H.M. Vazquez, R. He, T. Tan, Gabor ordinal measures for face recognition, IEEE Trans. Inf. Forensics Secur. 9 (1) (2014) 14–26.
[4] Z. Lei, M. Pietikainen, S.Z. Li, Learning discriminant face descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 289–302.

[5] J. Lu, V.E. Liong, X. Zhou, J. Zhou, Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (10) (2015) 2041–2056.
[6] W. Liu, J. Wang, R. Ji, Y. Jiang, S. Chang, Supervised hashing with kernels, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2074–2081.
[7] G. Lin, C. Shen, Q. Shi, A. van den Hengel, D. Suter, Fast supervised hashing with decision trees for high-dimensional data, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[8] R. He, Y. Cai, T. Tan, L.S. Davis, Learning predictable binary codes for face indexing, Pattern Recognit. 48 (10) (2015) 3160–3168.
[9] S. Stevens, On the theory of scales of measurement, Science 103 (2684) (1946) 677–680.
[10] Z. Sun, L. Wang, T. Tan, Ordinal feature selection for iris and palmprint recognition, IEEE Trans. Image Process. 23 (9) (2014) 3922–3934.
[11] Z. Sun, T. Tan, Ordinal measures for iris recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (12) (2009) 2211–2226.
[12] L. Xiao, R. He, Z. Sun, T. Tan, Coupled feature selection for cross-sensor iris recognition, in: Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems, 2013.
[13] Y.C. Chen, V. Patel, S. Shekhar, R. Chellappa, P. Phillips, Video-based face recognition via joint sparse representation, Automatic Face and Gesture Recognition, 2013.
[14] G. Zhang, R. He, L. Davis, Jointly learning dictionary and subspace structure for video-based face recognition, in: Proc. Asian Conference on Computer Vision, 2014.
[15] L. Xu, J. Neufeld, B. Larson, D. Schuurmans, Maximum margin clustering, in: Proc. Conference on Neural Information Processing Systems, 2004.
[16] T. Tan, X. Zhang, Z. Sun, H. Zhang, Noisy iris image matching by using multiple cues, Pattern Recognit. Lett. 33 (8) (2012) 970–977.
[17] N. Liu, M. Zhang, H. Li, Z. Sun, T. Tan, Deepiris: learning pairwise filter bank for heterogeneous iris verification, Pattern Recognit. Lett. 82 (2016) 154–161.
[18] S. Liao, Z. Lei, S.Z. Li, X. Yuan, R. He, Structured ordinal features for appearance-based object representation, Analysis and Modeling of Faces and Gestures, 2007.
[19] Z. Cao, Q. Yin, X. Tang, J. Sun, Face recognition with learning-based descriptor, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[20] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, PCANet: a simple deep learning baseline for image classification?, arXiv: 1404.3606v2 (2014) 1–16.
[21] Y. Weiss, A. Torralba, R. Fergus, Spectral hashing, in: Proc. Conference on Neural Information Processing Systems, 2009, pp. 1753–1760.
[22] B. Kulis, K. Grauman, Kernelized locality-sensitive hashing for scalable image search, in: Proc. IEEE International Conference on Computer Vision, 2009.
[23] W. Liu, C. Mu, S. Kumar, S.-F. Chang, Discrete graph hashing, in: Proc. Conference on Neural Information Processing Systems, 2014, pp. 3419–3427.
[24] J. Wang, S. Kumar, S.F. Chang, Semi-supervised hashing for scalable image retrieval, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2010.
[25] Y. Mu, G. Hua, W. Fan, S.-F. Chang, Hash-svm: scalable kernel machines for large-scale visual classification, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[26] F. Shen, C. Shen, W. Liu, H.T. Shen, Supervised discrete hashing, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[27] Y. Li, R. Wang, Z. Huang, S. Shan, X. Chen, Face video retrieval with image query via hashing across euclidean space and riemannian manifold, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[28] K. Grauman, R. Fergus, Learning binary hash codes for large-scale image search, Mach. Learn. Comput. Vision 411 (2013) 49–87.
[29] R. Xia, Y. Pan, H. Lai, C. Liu, S. Yan, Supervised hashing for image retrieval via image representation learning, in: Proc. AAAI Conference on Artificial Intelligence, 2014, pp. 2156–2162.
[30] J. Wang, H.T. Shen, J. Song, J. Ji, Hashing for similarity search: a survey, arXiv: 1408.2927 (2014).
[31] Z. Cui, H. Zhang, S. Lao, X. Chen, Image sets alignment for video-based face recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[32] C. Hu, X. Lu, M. Ye, W. Zeng, Singular value decomposition and local near neighbors for face recognition under varying illumination, Pattern Recognit. 64 (2017) 60–83.
[33] R. He, T. Tan, L. Wang, Robust recovery of corrupted low-rankmatrix by implicit regularizers, IEEE Trans. Inf. Forensics Secur. 36 (4) (2014) 770–783.
[34] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: a library for large linear classification, JMLR 9 (2008) 1871–1874.
[35] E. Grave, G. Obozinski, F. Bach, Trace lasso: a trace norm regularization for correlated designs, in: Proc. Conference on Neural Information Processing Systems, 2011.
[36] D. Cai, X. He, J. Han, Spectral regression for efficient regularized subspace learning, in: Proc. IEEE International Conference on Computer Vision, 2007, pp. 1–7.
[37] A. Hedayat, W.D. Wallis, Hadamard matrices and their applications, Ann. Stat. 6 (1978) 1184–1238.
[38] Y. Gong, S. Lazebnik, Iterative quantization: a procrustean approach to learning binary codes, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2011.
[39] W. Zhang, S. Shan, W. Gao, X. Chen, H. Zhang, Local gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition, in: Proc. IEEE International Conference on Computer Vision, 2005.
[40] Z. Lei, R. Chu, R. He, S. Liao, S.Z. Li, Face recognition by discriminant analysis with gabor tensor representation, in: IAPR/IEEE International Conference on Biometrics, 2007.
[41] R. Liu, S. Li, X. Yuan, R. He, Online determination of track loss using template inverse matching, International Workshop on Visual Surveillance, 2008.
[42] Y.L. Cun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Hand-written digit recognition with a back-propagation network, NIPS, 1990.
[43] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[44] X. Wu, R. He, Z. Sun, T. Tan, A light cnn for deep face representation with noisy labels, CoRR abs/1511.02683 (2015).
[45] S. Zhang, R. He, Z. Sun, T. Tan, Multi-task convnet for blind face inpainting with application to face verification, in: International Conference on Biometrics, 2016.
[46] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, Y. Ma, Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix, UIUC Technical Report, 2009.
[47] A. Gionis, P. Indyk, R. Motwani, Similarity search in high dimensions via hashing, VLDB, 1999.
[48] T. Kim, O. Arandjelovic, R. Cipolla, Discriminative learning and recognition of image set classes using canonical correlations, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007) 1005–1018.
[49] R. Wang, S.G. Shan, X.L. Chen, W. Gao, Manifold-manifold distance with application to face recognition based on image set, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[50] Y.H. amd A. Mian, R. Owens, Sparse approximated nearest points for image classification, in: Proc. IEEE International Conference on Computer Vision, 2011.
[51] R. Wang, H. Guo, L. Davis, Q. Dai, Covariance discriminative learning: A natural and efficient approach to image set classification, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[52] M. Yang, P. Zhu, L.V. Gool, L. Zhang, Face recognition based on regularized nearest points between image sets, Automatic Face and Gesture Recognition, 2013.
[53] E.G. Ortiz, A. Wright, M. Shah, Face recognition in movie trailers via mean sequence sparse representation-based classification, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[54] K. Lee, J. Ho, M. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2003.
[55] R. Gross, J. Shi, The CMU Motion of Body (MoBo) Database, Technical Report, Technical Report CMU-RI-TR-01-18, Robotics Inst., 2001.
[56] M. Kim, S. Kumar, V. Pavlovic, Rowley, Face tracking and recognition with visual constraints in real-world videos, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2008.

Ran He received the BE and MS degrees in computer science from Dalian University of Technology, and the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2001, 2004, and 2009, respectively. Since September 2010, he has been with the National Laboratory of Pattern Recognition, where he is currently an associate professor. He currently serves as an associate editor of Neurocomputing (Elsevier) and serves on the program committees of several conferences. His research interests include information theoretic learning, pattern recognition, and computer vision.

Tieniu Tan received the BSc degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and the MSc and PhD degrees in electronic engineering from Imperial College London, United Kingdom, in 1986 and 1989, respectively. He is currently a professor in the Center for Research on Intelligent Perception and Computing (CRIPAC), Natinal Laboratory of Pattern Recognition (NLPR) at the Institute of Automation, Chinese Academy of Sciences (CASIA). His current research interests include biometrics, image and video understanding, information hiding, and information forensics. He is a fellow of the IEEE and the IAPR (International Association of Pattern Recognition).

Larry S. Davis received the BA degree from Colgate University in 1970 and the MS and PhD degrees in computer science from the University of Maryland in 1974 and 1976, respectively. From 1977 to 1981, he was an assistant professor in the Department of Computer Science, University of Texas, Austin. He returned to the University of Maryland as an associate professor in 1981. From 1985 to 1994, he was the director of the University of Maryland Institute for Advanced Computer Studies. He is currently a professor at the Institute and in the Computer Science Department, as well as the chair of the Computer Science Department. He is known for his research in computer vision and high-performance computing. He has published more than 100 papers in journals and has supervised more than 20 PhD students. He is an associate editor of the International Journal of Computer Vision and an area editor for Computer Models for Image Processing: Image Understanding. He has served as the program or general chair for most of the field's major conferences and workshops, including the Fifth International Conference on Computer Vision, the 2004 Computer Vision and Pattern Recognition Conference, the 11th International Conference on Computer Vision. He became a fellow of the IEEE in 1997.

Zhenan Sun received the B.E. degree in industrial automation from the Dalian University of Technology, Dalian, China, in 1999, the M.S. degree in system engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2002, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2006. He joined the National Laboratory of Pattern Recognition, Center of Biometrics and Security Research, CASIA, as a Faculty Member in 2006. He is currently a Full Professor with CASIA. His current research interests include biometrics, pattern recognition, and computer vision.