PERSON RE-IDENTIFICATION USING MATRIX COMPLETION

Kai Liu[1], *Xin Guo*[1], *Zhicheng Zhao*[1,2], *Anni Cai*[1,2]

[1]School of Information and Communication Engineering
[2]Beijing Key Laboratory of Network System and Network Culture
Beijing University of Posts and Telecommunications, Beijing, China

ABSTRACT

Person re-identification is a challenging problem in multi-camera surveillance systems. In this paper, we formulate person re-identification as a cross-camera feature construction problem to overcome the feature variation between different camera spaces. The linear transformation of color information between probe and gallery camera spaces makes the stacked matrix, which concatenates features from these two camera spaces, rank deficient. From the feature observed in probe camera space we can construct its corresponding feature in gallery camera space by completing unknown entries on the relevant positions of the stacked matrix, and then match the constructed probe feature with features in gallery camera space. We also introduce additive noise term into the model to deal with the adverse effects caused by illumination variation with time. Experimental results demonstrate the proposed approach outperforms the metric learning methods as well as simple nearest neighbor search, and obtains a competitive performance compared with the state-of-the-art methods.

Index Terms— Person re-identification, matrix completion, rank minimization, linear relationship.

1. INTRODUCTION

Person re-identification(PRID) has attracted increasing attention in recent years due to its important role in multi-camera surveillance systems. It consists in recognizing individuals through images taken from two or more non-overlapping camera views. The main difficulties of PRID come from the changes in illumination, viewpoint, camera parameters, occlusions in different camera views and the low resolution of images. Usually, it is assumes that individuals wear the same clothes under different camera views[1], and thus appearance-based methods are widely employed in PRID. Among various features used in appearance-based methods, color is the most popular one [2, 1, 3, 4] due to its simplicity and effectiveness.

This work was supported by Chinese National Natural Science Foundation (90920001, 61101212), National High and Key Technology Research and Development Program (2012AA012505, 2012BAH63F00), National S&T Major Project of the Ministry of S&T (2012ZX03005008), the Special Funds of Beijing Municipal Co-construction Project, and the Fundamental Research Funds for the Central Universities.

According to the Lambertian reflectance model, when lighting conditions of two cameras both keep relatively stable, a linear transformation holds between color information of images taken by the two cameras for the same person. If we compose two matrices X and Y respectively with color features from two cameras for the same set of people, each column being the feature of one person, then the linear transformation T, *i.e.*$X = T^T Y$, will make the stacked matrix Z concatenated by $[X; Y]$ rank deficient.

Based on the above facts and inspired by rank minimization, we propose to formulate PRID as a matrix completion problem. Specifically, from the observed feature in probe camera space we can construct its corresponding feature in gallery camera space by completing unknown entries of the stacked matrix while minimizing its rank, and then match the constructed probe feature with gallery features in gallery camera space. Moreover, to cope with the uncontrolled illumination variations in real surveillance scenarios, we treat temporal variations of illumination as an additive noise term of the linear model, which makes the constructed feature immune to the interference of illumination variations.

Existing studies on PRID can be categorized into two families of methods. The first category views PRID as a matching problem, it seeks distinctive and robust feature descriptors and the distance between features is directly measured in spite of them coming from two different camera spaces, *e.g.*, SDALF [1]. The second category does not directly make feature comparison between two original camera spaces, but tries to seek a metric that describes the transition from the original feature space to a latent space. In the latent space, the distances between intra-class samples and those between inter-class samples become more discriminative, *e.g.*, Large Margin Nearest Neighbor (LMNN) [5] and Canonical Correlation Analysis (CCA).

Different from metric learning methods, our method doesn't need to pursuit the third latent space, but measures the feature distance in one of the two original camera spaces. Our method is also different from the simple matching methods, *e.g.* Nearest Neighbor search (NN), which ignores the difference of spaces where images lie in. Methodologically, our work is related with low rank matrix recovery [6, 7], which aims to recover original data with low rank structure. Sim-

ilar to this idea, Goldberg *et al.* [8] and Cabral *et al.* [9] focus on multi-label classification problem and predict labels through completing unknown entries in a matrix composed of instances and corresponding labels.

The main contributions of this work are two-folds. (1) We formulate the PRID as a matrix completion problem. To the best of our knowledge, it hasn't been investigated before. (2) We consider the color distortion attributed to illumination variation as additive noise component of linear model, which describes the color transformation across camera views.

2. PERSON RE-IDENTIFICATION USING MATRIX COMPLETION

2.1. Linear model of Color Transformation

According to the Lambertian reflectance model, in which the apparent brightness of an ideal diffusely reflecting surface to an observer is the same regardless of the observer's viewing angle, the image response of a digital camera imaging system depends on three physical variables: illumination spectrum and intensity,reflection property of the materials and camera sensor parameter.

For PRID, we make the following three assumptions: (1) clothes are matte, dull surfaces and clothes patches are locally planar; (2) clothes patches are illuminated with spatially uniform illumination, and illumination condition in each camera view keeps relatively stable with time; (3) narrow-band camera sensor, *i.e.*, the sensitivities of camera sensors are delta-functions.

Under such assumptions, the color change of a patch across cameras only depends on the difference of sensor parameters and illumination conditions between different camera views. Changes in illumination conditions can be modeled by a diagonal mapping base on von Kries model [10]. Thus for the same person, a linear transformation holds between color information of images from two different cameras because of the multiplicative combination of the diagonal mapping and approximately linear change in sensor response that can be assumed.

2.2. Formulation with Matrix Completion

In multi-camera surveillance systems, suppose a linear transformation $\mathcal{T}: \mathcal{X} \rightarrow \mathcal{Y}$ exists from one camera space $\mathcal{X}$ to another camera space $\mathcal{Y}$ in terms of color information. Denote the color feature vectors in $\mathcal{X}$ and $\mathcal{Y}$ by $x \in \mathbb{R}^F$ and $y \in \mathbb{R}^K$ respectively, where F and K denote the feature dimensions and F is equal to K for PRID. The linear transformation $T \in \mathbb{R}^{F \times K}$ satisfies the following equation:

$$x = T^T y. \tag{1}$$

Suppose we have N pairs of features $\{x_i, y_i\}, i \in \{1, \ldots, N\}$ extracted from images of two cameras, and then combine them as matrix Z with the following pattern:

$$Z = \begin{bmatrix} y_1, y_2, ..., y_N \\ x_1, x_2, ..., x_N \end{bmatrix}. \tag{2}$$

The stacked matrix Z should be rank deficient since the linear relationship between x and y holds.

By arranging samples of X and Y as in Z, one could easily construct features in one space from another space. More Specifically, suppose K pairs of features $\{x_i, y_i\}, i \in \{1, \ldots, K\}$ have already been known and $U = N - K$ features $\{x_j\}, j \in \{K+1, \ldots, N\}$ from one camera space are also observed. The features $\{y_j\}, j \in \{K+1, \ldots, N\}$ in another camera space could then be learned through minimizing the rank of matrix Z, *i.e.*,

$$\min_Z rank(Z) \quad s.t. \quad P_\Omega(Z) = Z_0, \tag{3}$$

where Z_0 is the matrix composed by known features with zeros at locations of unknown features and $P_\Omega(Z)$ denotes the entries in Z at the corresponding locations of known features in Z_0. The objective function forces the rank of the recovered matrix Z as low as possible, while the entries of Z at locations of the known features are kept equal to the observed ones through the constraint.

However, the assumptions for linear model of color information, which is described in section 2.1, may not be fully satisfied in real world due to various practical factors such as illumination variations with time. To cope with this problem, we define the corrupted observed matrix Z_0 as a sum of Z and an error term E, then Eq. 3 can be rewritten as follows:

$$\min_Z rank(Z) \quad s.t. \quad Z_0 = P_\Omega(Z) + E, \tag{4}$$

where Z_0,Z and E have the same dimensions and the constraint prevents large distortions of Z from observed entries in Z_0. We note that this problem is equivalent to

$$\min_Z \mu\, rank(Z) + \frac{1}{|P_\Omega(Z)|} \sum_{i \in P_\Omega(Z)} \ell(z_i, z_{0i}), \tag{5}$$

where $|\cdot|$ denotes element counting operator and $|P_\Omega(Z)|$ identifies the number of entries in matrix $P_\Omega(Z)$. $\ell(z_i, z_{0i})$ is the loss function between observed entry z_{0i} and recovered entry z_i, which is defined as the least square error:

$$\ell(z_i, z_{0i}) = \frac{1}{2}(z_i - z_{0i})^2. \tag{6}$$

The parameter μ is positive trade-off weights balancing the feature error correction and low rank of Z.

Since the rank minimization is hard to be solved, we relax Eq.5 to a nuclear norm minimization problem:

$$\min_Z \mu\, \|Z\|_* + \frac{1}{|P_\Omega(Z)|} \sum_{i \in P_\Omega(Z)} \ell(z_i, z_{0i}), \tag{7}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, *i.e.*,the sum of the singular values.

For a probe feature in probe camera space X, we can first construct its corresponding feature in gallery camera space Y using Eq.7, and then perform matching of the constructed probe feature with gallery features in gallery camera space Y, *i.e.*, the original cross-camera matching problem between space X and space Y now becomes matching problem in camera space Y. In this way, the PRID problem is converted to matrix completion with unknown or missing data. The latter has been well exploited in [6]. Denote y'_j as the constructed feature in gallery camera space from probe feature x_j, y'_j can then be classified as one of persons $\{K+1,\ldots,N\}$ by comparing the distances with feature set $\{y_u\}, u \in \{K+1,\ldots,N\}$ in gallery space Y:

$$min_u \ dist(y'_j, y_u), \ u = K+1,\ldots,N, \tag{8}$$

where $dist(\cdot,\cdot)$ denotes the Euclidean distance between two feature vectors. We should note that although the absolute distance between y_u and y'_j may be different from that between y_u and x_j in original space, the nearest neighbor search as Eq.8 guarantees the same matching result.

2.3. Optimization Procedure

The matrix completion problem in Eq.7 is a Nuclear Norm Minimization problem without constraints, which is convex and thus has a global optimal solution. Although the nuclear norm minimization can be reformulated to Semi-definite Program (SDP), current SDP solvers are not applicable to our problem due to the large dimension of Z. Alternatively, traditional fixed point approach has been devised to efficiently optimize this kind of problems. In this paper, we use Fixed Point Continuation (FPC) [11] method to solve our problem. This method consists of two alternating steps:

1. Gradient step: $A = f(Z) = Z - \tau g(Z)$ with step size τ and gradient $g(Z)$ is given by the loss function $\sum_{i \in P_\Omega(Z)} \ell(z_i, z_{0i})$ as follows:
$$g(z_i) = \begin{cases} \frac{1}{|P_\Omega(Z)|}(z_i - z_{0i}), & i \in P_\Omega(Z) \\ 0, & \text{otherwise} \end{cases}$$
2. Shrinkage step: $Z = S_{\tau\mu}(Z)$ which does SVD first and then applies soft shrink on the singular values to reduce the nuclear norm.

The complete FPC algorithm is summarized in Algorithm 1.

3. EXPERIMENTS

In this section, we show extensive experiments to evaluate our approach, providing comparisons with state-of-the-art methods on a synthetic dataset and VIPeR [12] dataset. On these two datasets, we compare our matrix completion approach with Nearest Neighbor search (NN) using Euclidean distance and classical metric learning methods such as LMNN and CCA. We also show comparisons of our approach to some state-of-the-art methods. To reduce the bias, experiments are run over 10 trials and average CMC curves[12] are reported.

Algorithm 1 FPC algorithm for solving Matrix Completion.

Input: Initial matrix Z_0, parameter μ, step size τ.
Initialize Z as the rank-1 approximate of Z_0.
for $\mu_1 > \mu_2 > \cdots > \mu_k = 10^{-5}$ **do**
  while Not converged **do**
    Gradient Descent: $A = f(Z) = Z - \tau g(Z)$
    Shrink: compute SVD of $A = U\Lambda V^T$
      compute $Z = U \max(\Lambda - \tau\mu, 0)V^T$
  end while
end for
Output: Complete Matrix Z.

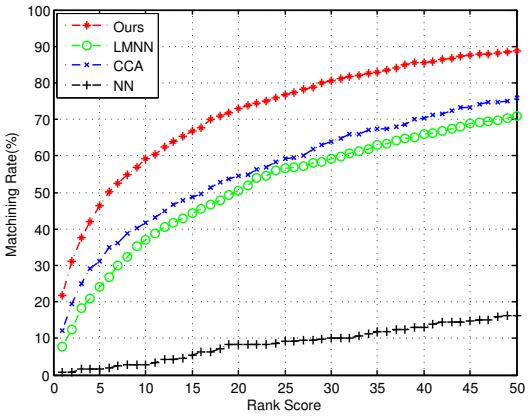

Fig. 1. Average CMC curves of our approach, LMNN, CCA, and NN on the synthesized dataset.

Similar to [11], we set parameter μ starting at $\mu_1 = 0.25\sigma_1$, where σ_1 is the largest singular value of matrix Z (the unknown entries set to 0), and decreasing by $\mu_k = 0.25\mu_{k-1}$ until 10^{-8}. The step size of gradient descent is set as $\tau = |P_\Omega(Z)|$. Convergence is defined as a relative change in objective function smaller than 10^{-2}.

3.1. Experiments on Synthetic Dataset

To construct the synthetic dataset, we first generate matrix $Y \in \mathbb{R}^{D\times N}$ and transformation matrix $T \in \mathbb{R}^{D\times D}$, in which the value of each entry is randomly chosen from the range of [0,1]. Then matrix $X \in \mathbb{R}^{D\times N}$ is derived by $X = T^T Y$ to ensure the linear relationship between X and Y. Each column in matrices X and Y represents a sample with dimension D. At last, the Gaussian noise with zero mean and variance 0.3 is added to X and Y to simulate the error in real world. In our experiments, D and N are set as 50 and 632 respectively.

From the synthetic dataset, we randomly select 316 pairs from X and Y to form the test set, and the rest are used for training. In the test set, the samples from space Y compose gallery set and the samples from space X compose the probe set. Experimental results are illustrated in Fig. 1.

As can be seen in Fig. 1, when data between two different spaces follow linear relationship, our approach and metric learning based methods obtain much better performances than Nearest Neighbor search method based on Euclidean dis-

Fig. 2. Example images from the VIPeR database. Each column represents the matched pair of the same person, and upper and lower row correspond to different appearances from camera X and camera Y.

tance. Due to the introduction of the error term into the linear model, our approach can cope with the interference of real-world error and perform better than LMNN and CCA.

3.2. Experiments on VIPeR dataset

VIPeR[1] is a widely used dataset in PRID. It consists of 632 pedestrian image pairs taken from two different cameras. Each pedestrian image pair has been taken from arbitrary viewpoints under varying illumination conditions. Representative samples are shown in Fig. 2.

Due to the simplicity and effectiveness of color information, our algorithm uses color histogram as image feature. Specifically, for each image, we normalize it to 200×100 pixels and divide it into ten horizontal stripes. For each stripe, the RGB,HS and YCbCr features are extracted and represented as a histogram. Each person image is thus represented by a feature vector in the 1216 dimensional feature space.

Exactly following the setting in [4], we randomly select 316 image pairs to form the test set, and the rest are used for training. In the test set, the gallery set consists of one image for each person, and the remaining images are used as the probe set. For performance comparisons between different methods, the same color features are used in all these four methods without any other supplementary information. The obtained CMC curves on VIPeR dataset are shown in Fig. 3.

Due to the changes in illumination, pedestrian pose and etc., the color features from two viewing conditions may not have simple correspondence between them, so the non-learning based NN method gets the worst performance. In contrast, three learning based methods show more satisfactory results. Matrix completion method utilizes the linear model of color transformation across cameras and introduces additive noise term into the model to cope adverse effects caused by illumination variations with time, so the proposed method achieves better performance than classical metric learning methods, LMNN and CCA.

Moreover, we show the comparisons of our approach to the state-of-the-art PRID methods in Table 1. As can be seen from the table, our method outperforms the state-of-the-art

[1]The dataset is available at http://vision.soe.ucsc.edu/?q=node/178/

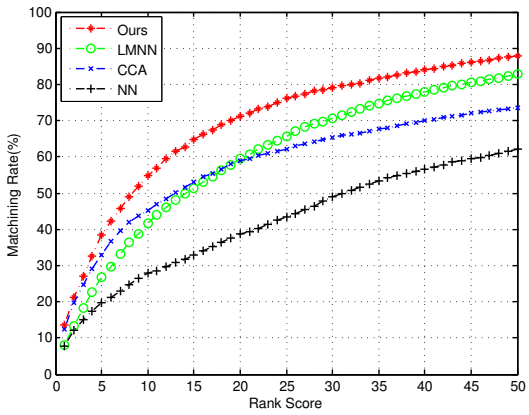

Fig. 3. Average CMC curves of our approach, LMNN, CCA, and NN on the VIPeR dataset.

Methods	Rank=1	10	20	50
ELF [4]	12	43	60	81
PRSVM [3]	13	50	67	85
SDALF [1]	**20**	53	67	84
PRDC [2]	16	53.8	70	87
Ours	12.7	**56.1**	**72**	**88**

Table 1. Comparisons of matching rates (%) at different ranks on the VIPeR dataset. Results from four state-of-the-art methods are listed for comparison with our approach.

PRID methods in most rank configurations. Moreover, we should note that SDALF utilizes color features (wHSV, M-SCR) and texture feature(RHSP), while ELF, PRSVM and PRDC all use 8 color channels (RGB, HS and YCbCr) and 2 texture filters (Gabor and Schmid) to represent images. In contrast, our method only uses color features for image representation. Confronting the complicated conditions in viewpoint, background, occlusions and low resolution of images, color features are more stable than texture features, making our method more reliable for real-world applications.

4. CONCLUSION

In this paper, we focus on person re-identification in multi-camera surveillance systems and cast it as a matrix completion problem. By modeling the linear relationship of color transformation between different cameras, our approach can construct the corresponding feature in gallery camera space from the observed feature in probe camera space by completing unknown entries on relevant positions of data matrix. We also consider the color distortion as an additive noise component in linear model and this trick makes the constructed image descriptors immune to interference of illumination variation with time. Experimental results on a synthetic dataset and a benchmark dataset demonstrate that our approach outperforms metric learning methods as well as simple nearest neighbor search and achieves competitive performance compared with the state-of-the-art methods.

5. REFERENCES

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR'10*.

[2] W.S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR'11*.

[3] B. Prosser, W.S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking," in *BMVC'10*.

[4] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV'08*.

[5] K.Q. Weinberger, J. Blitzer, and L.K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *NIPS'06*.

[6] E.J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, 2009.

[7] J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *NIPS'09*.

[8] A.B. Goldberg, X. Zhu, B. Recht, J.M. Xu, and R. Nowak, "Transduction with matrix completion: Three birds with one stone," in *NIPS'10*.

[9] R.S. Cabral, F. De la Torre, J.P. Costeira, and A. Bernardino, "Matrix completion for multi-label image classification," in *NIPS'11*.

[10] J. Von Kries, "Influence of adaptation on the effects produced by luminous stimuli," 1970.

[11] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, 2011.

[12] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Workshop on PETS'07*.