EXPLOITING LOW-RANK STRUCTURES FROM CROSS-CAMERA IMAGES FOR ROBUST PERSON RE-IDENTIFICATION

Ming-Hang Fu\textsuperscript{1}, Yu-Chiang Frank Wang\textsuperscript{2}, Chu-Song Chen\textsuperscript{1}\\
\textsuperscript{1}Institute of Information Science, Academia Sinica, Taipei, Taiwan \textsuperscript{2}Research Center for IT Innovation, Academia Sinica, Taipei, Taiwan

ABSTRACT

Matching individuals across non-overlapping camera views is known as the problem of person re-identification. In addition to significant visual appearance variations due to lighting, view angle, etc. changes, one might encounter corrupted data due to background clutter and occlusion, or even missing data at some camera views in practical scenarios. To address the above challenges, we present a novel approach to robust person re-identification, particularly aiming at handling missing and corrupted image data across camera views. Based on the technique of low-rank matrix decomposition, our proposed algorithm observes the low-rank structure of cross-view data, which is able to disregard extreme/sparse errors while the missing instances can be recovered automatically. Our experiments will confirm the effectiveness and robustness of our method, which is shown to outperform several baseline and state-of-the-art person re-identification approaches.

Index Terms— Person Re-Identification, Low-Rank Matrix Recovery

1. INTRODUCTION

Automatic matching or recognizing individuals across different camera views is known as the problem of person re-identification. Typically, one needs to deal with significant variations like viewpoint, lighting or other visual appearances across different cameras (see Figure 1 for example). This is the reason why person re-identification has been a challenging task for researchers and engineers in related fields.

In order to handle the aforementioned problems, several visual features have been proposed for representing the images of different camera views. For example, Gray and Tao [1] utilized a set of localized visual features for describing cross-camera images, so that additional robustness would be introduced. Bak \textit{et al.} considered Haar and DCD based descriptors [2], and also considered the extraction of spatial covariance regions for improved invariance capabilities [3]. On the other hand, Bazzani \textit{et al.} [4] exploited chromatic information and constructed epitome-based histogram features for cross-view image representation.

Recent research attention has been directed to the development of learning-based algorithms, with the goal of modeling the differences in visual appearance across camera views. For example, Bäuml \textit{et al.} [6] applied multinomial logistic regression for minimizing matching and reconstruction losses. In order to improve the matching accuracy, Prosser \textit{et al.} [7] utilized rankSVM for converting the person re-identification problem into a ranking problem. On the other hand, Zheng \textit{et al.} [8] presented a Probabilistic Relative Distance Comparison (PRDC) model to learn the optimal distances between images across different views. Tao \textit{et al.} [9] advanced metric learning and introduced RS-KISS for improved performance. Some works regarded person re-identification problems as cross-domain problems. For example, An \textit{et al.} [10] used Regularized Canonical Correlation Analysis (RCCA) to project cross-camera data into a subspace for recognition, and Avraham \textit{et al.} [11] proposed Implicit Camera Transfer (ICT) for associating cross-camera data for re-identification purposes.

The technique of matrix completion was advanced in [12], which views test data at one view as the inputs, while those to be predicted at the other as missing data. However, since they assumed that training data across camera views can be collected during the learning process, it might not be preferable to extend their approach to real-world scenarios in which the training data might not be available at either view. In addition, most prior learning-based approaches were not explicitly designed to handle both missing and outlier data during training. As a result, not only possible missing data, it

Fig. 1. Example images from the VIPeR dataset [5]. Each column shows images of the same person at two different camera views. Note that background clutter or occlusion can be observed.
is practical to expect that both training and test image data across camera views might be partially occluded (due to extreme lighting conditions, clothing accessories, background regions, etc.). Based on the above observations, we propose a novel approach for person re-identification based on low rank matrix recovery (LR) [13]. Inspired by [12], we exploit the low-rank structures from cross-camera images using the observed visual features. Our proposed formulation not only handles missing image data, it is also robust to corrupted image regions due to background clutter or occlusion. As a result, we are able to recover and match the image at the view of interest for re-identification purposes.

2. OUR PROPOSED METHOD

2.1. Color-Based Feature Representation

Color-based features have been widely applied for representing images across different camera views because of their robustness [1, 2, 3, 4]. However, due to lighting, etc. condition changes, the color features of different camera views might still exhibit significant variations. Therefore, one cannot expect the direct use of such features for recognition purposes.

As pointed out in [14, 15, 12], one can typically observe a linear relationship between the color features across different camera views. Let \( X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{m \times n} \) and \( Y = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^{m \times n} \) as a set of image pairs, in which \( x_i \) and \( y_i \) are the \( m \)-dimensional color features of the \( i \)th instance observed by cameras 1 and 2, respectively. According to [14, 12], a linear transform exists for converting \( X \) into \( Y \), even \( X \) and \( Y \) are captured under very different lighting conditions.

In our work, we consider the color features suggested by [12]. We divide each image into 16 overlapping horizontal stripes. For each stripe, the color histograms of RGB, HSV, and \( C_b, C_r \) are calculated (with 8 bins for each channel). As a result, a \((16 + 15) \times 8 \times 8 = 1984\) dimensional feature will be produced for representing an image at each camera view.

2.2. Exploiting Low-Rank Structures from Noisy and Missing Images Across Cameras

For the training stage of our proposed person re-identification method, a cross-view image matrix of \( Z = [X; Y] \in \mathbb{R}^{2m \times n} \) can be constructed, in which \( X \) and \( Y \) indicate \( n \) training image pairs observed by two non-overlapping cameras. It is worth noting that, we expect missing training data at either camera view (i.e., either \( x_i \) or \( y_i \) might not be available for instance \( i \)).

Assuming that \( X \) and \( Y \) are i.i.d., we expect to restore the missing instances by minimizing the rank of the matrix \( Z \) (i.e., \( \text{rank}(Z) \) will be between \( 2m \) and \( n \)). This is because that, as suggested in [14, 12], there exists some linear relationships between \( X \) and \( Y \). Thus, based on the technique of matrix completion, we aim at deriving \( A \in \mathbb{R}^{2m \times n} \) as a low-rank approximation version of \( Z \) by:

\[
\min_A \text{rank}(A) \quad \text{s.t.} \quad \begin{cases} A(i, 1 : m) = x_i, & \text{if } x_i \in \Omega, \\ A(i, m + 1 : 2m) = y_i, & \text{if } y_i \in \Omega, \end{cases}
\]

(1)

where \( \Omega \) indicates the set of observed instances (i.e., not missing data). We note that, the above formulation can be further written as follows:

\[
\min_A \text{rank}(A) \quad \text{s.t.} \quad Z = A + A_{\Omega},
\]

(2)

where

\[
A_{\Omega}(i, 1 : m) = \begin{cases} 0, & x_i \in \Omega, \\ -A(i, 1 : m), & \text{otherwise}, \end{cases}
\]

\[
A_{\Omega}(i, m + 1 : 2m) = \begin{cases} 0, & y_i \in \Omega, \\ -A(i, m + 1 : 2m), & \text{otherwise}. \end{cases}
\]

According to (2), if we have missing instances at either view in \( Z \), the corresponding and complementary vectors/features will be recovered in \( A \) and \( A_{\Omega} \), respectively.

For practical scenarios, we not only need to handle missing data, the observed image data might also be corrupted/occluded due to possible background or clutter regions. If such artifacts are presented, the solution to the above low-rank matrix recovery problem might not be preferable. To address this concern, we view such corrupted image features as outlier or noisy data, and propose a novel low-rank matrix decomposition algorithm for handling outlier \( E_{\Omega} \) and missing data \( A_{\Omega} \) (as depicted in Figure 2):

\[
\min_{A, E_{\Omega}} \text{rank}(A) + \lambda \| E_{\Omega} \|_0 \quad \text{s.t.} \quad Z = A + A_{\Omega} + E_{\Omega},
\]

(3)

where \( E_{\Omega_{ij}} = 0 \) if \((i, j) \notin \Omega\). This means that, if the \( i \)th instance has missing data at either camera view, the corresponding \( E_{\Omega_{ij}} \) will be zeroes. Otherwise, we aim at disregarding the corrupted image regions by minimizing the zero norm of \( E_{\Omega_{ij}} \) in (3).

To make the proposed problem more tractable [13], we solve the convex relaxation version of (3) instead:

\[
\min_{A, E_{\Omega}} \| A \|_* + \lambda \| E_{\Omega} \|_1 \quad \text{s.t.} \quad Z = A + A_{\Omega} + E_{\Omega},
\]

(4)

where \( \| A \|_* \) denotes the nuclear norm of \( A \), and \( \| E_{\Omega} \|_1 \) sums up the absolute values of each entry in \( E_{\Omega} \).
2.3. Optimization

To solve the proposed problem of (4), we apply the technique of Inexact ALM Method [13] and reformulate (4) as follow:

\[
\min_{A, A_\Omega, E_\Omega} F(A, A_\Omega, E_\Omega, L, \mu) = \min_{A, A_\Omega, E_\Omega} \| A \|_* + \lambda \| E_\Omega \|_1 + \mu L \nonumber
\]

\[
+ L, Z - A - A_\Omega - E_\Omega > + \mu \| Z - A - A_\Omega - E_\Omega \|_F^2 \nonumber
\]

in which \( L \in \mathbb{R}^{m \times n} \) is the Lagrange multiplier, while the equality constraint is regularized by parameter \( \mu \). To solve (5), we apply alternative optimization techniques for deriving the optimal \( A_\Omega, A_{\Omega k} \) and \( E_\Omega \). The pseudo code of our proposed algorithm is shown in Algorithm 1. We now discuss how we update the above variables in each iteration during the optimization process.

Updating \( A \)

To update \( A \), we fix \( A_{\Omega k}, E_\Omega \) and \( L \) in (5) for determining the optimal value of \( A \). In other words, at the \( k \)-th iteration, we update \( A_{k+1} \) by solving the following problem:

\[
A_{k+1} = \arg \min_{A} \| A \|_* + \lambda \| E_\Omega \|_1 + \mu \| Z - A - A_{\Omega k} - E_{\Omega k} \|_F^2 \nonumber
\]

\[
+ \mu \| Z - A - A_{\Omega k} - E_{\Omega k} \|_F^2 \nonumber
\]

\[
= \arg \min_{A} \frac{1}{\mu_k} \| A \|_* + \frac{1}{2} \| A - C_A \|_F^2. \tag{6}
\]

where \( C_A = Z - A_{\Omega k} - E_{\Omega k} + \frac{L_k}{\mu_k} \). As suggested by [16], the solution of the above problem can be solved as \( A_{k+1} = U_{\epsilon}(S)^{V^{T}} \), where \( [U, S, V] = SVD(C_A) \) and \( \epsilon = \mu_k^{-1} \). Note that \( U_{\epsilon}(S)^{V^{T}} \) is defined by element-wise \( \epsilon \) thresholding of \( S \) (see [16] for more details).

Updating \( E_\Omega \)

To calculate the optimal \( E_\Omega \) in (4), we fix \( A, A_{\Omega k} \) and \( L \) and reformulate (5) as follows:

\[
E_\Omega = \arg \min_{E_\Omega} \lambda \| E_\Omega \|_1 + \frac{\mu_k}{2} \| Z - A_{k+1} - A_{\Omega k} - E_{\Omega k} \|_F^2 \nonumber
\]

\[
= \arg \min_{E_\Omega} \frac{1}{\mu_k} \| \bar{S} \|_1 + \frac{1}{2} \| \bar{S} - C_{E_\Omega} \|_F^2, \tag{7}
\]

where \( C_{E_\Omega} = Z - A_{k+1} - A_{\Omega k} + \frac{L_k}{\mu_k} \). The above optimization problem can be solved by \( \ell_1 \)-minimization techniques (e.g., [17]), and thus the solution can be derived as \( E_{\Omega k+1} = T_\epsilon(C_{E_\Omega}) \) with \( \epsilon = \lambda \mu_k^{-1} \).

Updating \( A_{\Omega} \)

Finally, we fix \( A, E_\Omega \) and \( L \) in (5) for updating \( A_{\Omega} \) by:

\[
A_{\Omega} = \arg \min_{A_{\Omega} E_\Omega} \| L_k, Z - A_{k+1} - A_{\Omega k} - E_{\Omega k} \|_F^2 \nonumber
\]

\[
= \arg \min_{A_{\Omega} E_\Omega} \| L_k, Z - A_{k+1} - A_{\Omega k} - E_{\Omega k} \|_F^2 \nonumber
\]

\[
= \frac{1}{\mu_k} \| Z - A_{k+1} - A_{\Omega k} - E_{\Omega k} \|_F^2 + \frac{L_k}{\mu_k} \| Z - A_{k+1} - A_{\Omega k} - E_{\Omega k} \|_F^2 \nonumber
\]

\[
= \frac{1}{\mu_k} \| Z - A_{k+1} - A_{\Omega k} - E_{\Omega k} \|_F^2 + \frac{L_k}{\mu_k} \| Z - A_{k+1} - A_{\Omega k} - E_{\Omega k} \|_F^2 \tag{8}
\]

Once \( A_\Omega, E_\Omega \), and \( A_{\Omega k} \) are obtained, the matrix (i.e., the Lagrange multiplier) \( L \) can be simply updated by the last equation in Algorithm 1. The convergence of these matrices indicates the termination of our optimization process.

3. EXPERIMENTS

To evaluate the performance of our proposed method, we consider the the ViPeR dataset [5]. This dataset contains images of 632 subjects, while each has a pair of images captured by two non-overlapping cameras 1 and 2. We randomly and equally divide this dataset into two subsets (one for training and the other for testing), and repeat the process 10 times for reporting the average performance.

For the test data, we consider all instances observed by camera 1 as the probe set, while those at camera 2 as the gallery set to be recognized (as did in [12]). When performing re-identification, the matrix \( Z \) not only consists of 316 cross-view training instance pairs, it also contains 316 probe instances observed by camera 1 (the test instances at camera 2 are viewed as missing data). After the optimization of (4) is complete, the test instances at camera 2 will be recovered, and thus matching between such instances and the gallery set can be performed accordingly.

3.1. Re-Identification without Missing Data

We first consider the scenario in which the training data only exhibits occluded image regions. Thus, all 316 cross-view training instances are available during the derivation of \( A \)
Table 1. Recognition performance of different methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Rank=1</th>
<th>10</th>
<th>20</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>0.3</td>
<td>2.1</td>
<td>4.9</td>
<td>15.6</td>
</tr>
<tr>
<td>ELF [1]</td>
<td>12</td>
<td>43</td>
<td>60</td>
<td>81</td>
</tr>
<tr>
<td>PRSVM [7]</td>
<td>13</td>
<td>50</td>
<td>67</td>
<td>85</td>
</tr>
<tr>
<td>SDALF [18]</td>
<td>20</td>
<td>53</td>
<td>67</td>
<td>84</td>
</tr>
<tr>
<td>PRDC [8]</td>
<td>16</td>
<td>53.8</td>
<td>70</td>
<td>87</td>
</tr>
<tr>
<td>MC [12]</td>
<td>12.7</td>
<td>56.1</td>
<td>72</td>
<td>88</td>
</tr>
</tbody>
</table>
| Ours    | 16.4  | 57.5| 73.4| 88.4 (%)

in (4). Once the probe instances at camera 2 are recovered for performing re-identification, we apply the metric of the Bhattacharyya distance (as did in [18]) for matching such images with the gallery ones. In our experiments, the performance is carried out by the Cumulative Matching Characteristic (CMC), which indicates the recognition accuracy of finding the correct matches in the top rank-n candidates.

To compare our proposed method with others, we consider several state-of-the-art re-identification approaches in Table 1: ensemble of localized features (ELF) [1], person re-identification SVM (PRSVM) [7], symmetry-driven accumulation of local features (SDALF) [18], probabilistic relative distance comparison (PRDC) [8], and matrix completion (MC) [12]. We also consider nearest neighbor (NN), which can be viewed as a baseline approach for directly performing matching of probe and gallery images using our color features. From this table, it can be seen that our approach generally outperformed others. It is worth noting that, instead of extracting more sophisticated features like ELF and SDALF did, we only apply existing and standard color features for producing satisfactory performance.

3.2. Re-Identification with Corrupted and Missing Data

We now consider a more challenging and practical setting for evaluation. In addition to having the training data corrupted by background or clutter regions, we further randomly disregard 20% of the training instances at either camera view. As a result, one needs to solve person re-identification problems using missing and corrupted training cross-view data.

Based on the above setting, 20% of the subjects in the training set (i.e., 64 people) have only one image available. As for the test set, we again have all instances at camera 1 as the probe set while those at camera 2 as the gallery ones. We perform CMC comparisons with nearest neighbor (NN) and MC [12]. Figure 3 presents the recognition performance over different rank numbers. We see that, due to the capability of handling both corrupted and missing image data, our approach achieved the highest CMC results across different ranks, while our improvements over MC were more significant than those in Section 3.1.

We show example results in Figure 4 for verifying our robustness to corrupted image regions. It can be seen that, while background or clutter regions (e.g., bags) occlude parts of the images, our method successfully identified such regions (shown in red rectangles) when deriving the low-rank cross-view model in (4). This again supports the use of our proposed method for robust person re-identification.

4. CONCLUSION

We proposed a novel low-rank model for representing and associating cross-camera images. In particular, our model is able to handle corrupted image regions which result in sparse errors, and to recover missing data at either view. As a result, the problem of robust person re-identification can be solved. Our experiments confirmed that our approach outperformed state-of-the-art re-identification methods, which either required the extraction of sophisticated features or lacked the ability of handling corrupted/missing data. Thus, the effectiveness and robustness of our method can be verified.

Acknowledgement This work is supported in part by the National Science Council of Taiwan via NSC102-3111-Y-001-015, NSC102-2221-E-001-005-MY2, and NSC 101-2221-E-001-015-MY2.
5. REFERENCES


