A FEATURE FUSION STRATEGY FOR PERSON RE-IDENTIFICATION

Mu Gao¹, Haizhou Ai¹, Bo Bai²

¹Department of Computer Science and Technology, Tsinghua University, Beijing, China.,
²Huawei Technologies, Beijing, China.
gao13@mails.tsinghua.edu.cn, ahz@mail.tsinghua.edu.cn, baibo3@huawei.com

ABSTRACT

The problem of person re-identification, identifying the same person appeared in different camera views, is an important and challenging task in computer vision that has high potential application in areas like visual surveillance. In this paper we introduce a new feature fusion strategy for person re-identification that combines low-level Weighted Histograms of Overlapping Stripes (WHOS) features with mid-level color name descriptors and we adopt KISSME algorithm for person matching. Experiments on several public person re-identification datasets (VIPeR, i-LIDS and CA VIAR4REID) demonstrate that our approach achieves much better results compared with other state-of-the-art approaches.

Index Terms— Person re-identification, feature fusion, metric learning

1. INTRODUCTION

Person re-identification is the task of recognizing the same person over different camera views, which is an important and open problem in computer vision that has high potential application in areas like visual surveillance. Due to the influence of occlusions, pose changes, light conditions and low resolution, person re-identification under uncontrolled complex environments remains a challenging problem and gain more and more research interests in recent years [3-4, 7-10, 15-25].

The task of person re-identification can be mainly divided into two stages:

(1) Feature representation. Many methods have been proposed to obtain an efficient and robust feature expression for person re-identification and among them color has been proved to play an important role and is the most commonly used appearance feature [3, 4, 9-10, 15-18]. Except color histograms in different color spaces such as RGB, HSV, YCbCr, color name distribution is another efficient descriptor which is of special practical interest because of its effectiveness in recent researches. Weijer [1-2] proposed a new approach to learn color names from real-world images which mapped RGB values to 11 pre-defined colors (black, blue, brown, grey, green, orange, pink, purple, red, white and yellow). Similarly, Yang et al [3] proposed another mapping model between RGB values and 16 salient color names and applied it to person re-identification, of which the probability for each RGB value over certain color names is pre-calculated. Color name models can be seen as a new mid-level color descriptor in form of a probability distribution over certain number of color names. Although color is effective, it is not the only efficient feature and other features such as LBP texture descriptor and HOG feature can also be combined with color to strength its representation ability.

(2) Similarity computation. Direct distance such as Euclidean distance performs poorly in person re-identification. Metric learning, whose aim is to learn a Mahalanobis metric based on learning objective that keeps pairs of the same person closer and pairs of different persons far, has been proved successful in this task [3, 5, 11-14]. For instance, KISSME [5] is a simple yet effective distance metric learning algorithm based on statistical inference and has been widely used. Other metric learning methods include Large Margin Nearest
Neighbor Learning (LMNN) [11-12], Information Theoretic Metric Learning (ITML) [13] and Logistic Discriminant Metric Learning (LDML) [14].

Some pre-processing steps in person re-identification include but not limit to person segmentation and body part subdivision [3, 7] which aim to eliminate the influence of background and take local features of human body into account. Some post-processing steps are also proposed. Commonly used method is to regard it as a ranking or searching problem and adopt post-rank or searching techniques [8, 15, 21].

Our main contributions in this paper include: We propose a novel feature fusion strategy that combines high-dimensional low-level features (WHOS which combine color histograms and HOG) and low-dimensional mid-level semantic descriptors (color names distribution) and based on simple metric learning algorithm (KISSME) we achieve state-of-the-art results on several public datasets (VIPeR, iLDS, CAVIA4REID). The paper is organized as follows. In Section 2, the framework and algorithms of our approach are explained in detail. Section 3 describes the experiments and results on public datasets. The conclusion summarizes the contents of this paper.

2. OUR APPROACH

2.1. Pre-processing

Our pre-processing procedures include two main steps: person segmentation and body part subdivision.

Inspired by the work [6], we develop a lightweight variant of that method and found that it is very effective in doing person segmentation in various datasets. We made a little adaptation of the method i.e. dividing original 8 body parts parsing to only 2 parts parsing i.e. foreground and background. The masks after background removal are then used on original images to generate pedestrian silhouettes. The results on various images are found to be better compared with prior knowledge based methods and other existing methods.

The part-based body models used in current appearance descriptors can be divided into three categories: fixed, adaptive and learned models [7]. Fixed models, in which the sizes and positions of body parts are chosen a-priori, are simple yet proven to be effective in many pedestrian related tasks. We utilized a fixed part model in our approach that divides each pedestrian image into six horizontal stripes of equal size.

2.2. Feature representation

2.2.1. Global color name descriptor

In order to build a part-based representation, each pedestrian \( P \) is defined as a sequence of \( m \) stripes (here \( m \) is 6 and each stripe is equal size in our representation):

\[
P = [P_1, ..., P_m]
\]

Our global color name descriptor is only calculated in the person area in order to eliminate the influence of background noise. For each stripe, we extract color name features, serving as mid-level descriptors. The feature vector of the stripe \( j \) is denoted as:

\[
P_j = [CN_{1j}, ..., CN_{sj}]
\]

where \( CN_s \) is the color name probability for color \( s \).

We calculate the color name distribution for each stripe using the color name model in [1, 2]. The core of the model is to build a relationship between a RGB triplet and a certain probability distribution over 11 predefined colors. Actually in the model the RGB color space is discretized into \( 32 \times 32 \times 32 = 32768 \) indexes so it comes in the form of a \( 32768 \times 11 \) dimension lookup table. For the stripe \( j \) \((j = 1, 2, 3, 4, 5, 6 \) in our representation), the color name distribution \( CN_s \) is defined as:

\[
CN_{sj} = \frac{1}{N} \sum_{x_{RGB} \in R_j} p(CN_s|x_{RGB})
\]

where \( p(CN_s|x_{RGB}) \) is the probability of a certain pixel’s RGB value \((x_R, x_G, x_B)\) being assigned to a specific color name \( CN_s \). \( R_j \) is the foreground region of the stripe \( j \). \( N \) denotes the summation of pixels that are in \( R_j \) which is used for normalization and it is easy to see that \( \sum_{s=1}^{11} CN_{sj} = 1 \).

Then the global color name descriptor can be obtained as a 66-dimensional feature vector by concatenating each stripe’s CN features. We then use PCA to reduce the dimension to \( m \) (for different datasets we use different number because the number of persons varies in different datasets).

2.2.2. Combined feature representation

For low-level features, we choose to take color histograms in different color spaces, LBP and HOG into account. Lisanti et al [4] designed a descriptor namely Weighted Histograms of Overlapping Stripes (WHOS) that combines the main features stated above. So in order to have a comparison we decide to use the original WHOS descriptor as our low-level features. WHOS descriptor contains HS, RGB and HOG features, more implementation details can be found in [4].

\[
WHOS = [HS, RGB, HOG]
\]

The dimension of WHOS is 2960 in original representation, we use PCA to reduce the dimension to \( n \) (for different datasets we use different number because person number varies in different datasets) because the original dimension may be too high.

Our final representation then combines global color name descriptor and low-level WHOS descriptor in a fusion strategy.

\[
F = [CN, WHOS]
\]
Our combined feature representation is then a \((m+n)\) dimensional (varies in different datasets) feature vector by fusing mid-level CN and low-level WHOS. WHOS are high-dimension low-level features. HS and RGB are part-based color histograms in different color spaces that can compensate each other against illumination changes. HOG is a global feature which is proven to be effective in pedestrian related task. On the other hand, color name descriptor is a low-dimension mid-level descriptor which has more discriminative power. We combine them together in order to be more representative and robust. After fusion PCA is adopted to reduce the dimension if necessary.

2.3. Learning algorithms

We chose to use learning based methods in person matching step. Given two person samples \(x_i\) and \(x_j\), the Mahalanobis distance between them is

\[
d^2_{M}(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j)
\]

where \(M\) is a positive semi-definite matrix that metric learning methods aim to learn. In our approach we choose the KISSME algorithm. KISSME defines \(M\) as follows:

\[
M = \sum_S^{-1} - \sum_D^{-1}
\]

where

\[
\sum_S = \frac{1}{|S|} \sum_{x_i, x_j \in S} (x_i - x_j)(x_i - x_j)^T
\]

\[
\sum_D = \frac{1}{|D|} \sum_{x_i, x_j \in D} (x_i - x_j)(x_i - x_j)^T
\]

where \(S\) means similar pairs and \(D\) means dissimilar pairs respectively. More details of KISSME can be found in [5]. We keep the same settings in the algorithm except features and their dimension.

3. EXPERIMENTS

Our method is evaluated on three public datasets: VIPeR [27], i-LIDS [28] and CAVIA4REID [20]. In all our experiments we adopt the single-versus-single (SvsS) modality [4] which means that we have a single sample for each person in the probe set and also a single sample for each person in the gallery set. The approaches used for comparison are also evaluated in SvsS modality. The results are evaluated in form of top-k ranking accuracy and CMC (Cumulated Matching Characteristic) curve. Other settings are the same as stated in [5] such as half image pairs are randomly chosen for training and the remaining half pairs are used for test.

3.1. VIPeR

The first dataset we used for evaluating our approach is VIPeR, one of the most commonly used and challenging datasets for person re-identification, containing 1264 images of 632 different pedestrians. Each person has two images taken by two cameras from different view degree and the number of person used for test is 316. The images are scaled to be 128x48 pixels. Some examples are shown in Fig.2(a).

In this experiment the dimension of CN \(m\) are set to be 34 and the dimension of WHOS \(n\) are set to be 70 so the final feature dimension is 104. The results are shown in Table 1 and Fig.3(a). We can see that when used alone CN and WHOS along with KISSME can only achieve little accuracy improvement respectively but when fused as in our strategy it can achieve surprisingly very accurate results that is of a great improvement over other methods. The top-1 rank accuracy of our fusion strategy is 95.25% which is the best result as far as we know and greatly outperforms other approaches.

3.2. i-LIDS

The i-LIDS dataset contains person images in an airport taken from multiple camera views and is extracted from the i-LIDS multi-camera tracking scenario or i-LIDS MCTS dataset which is widely used for tracking evaluation purposes. It contains 476 images of 119 different persons. All images are normalized to 128x64 pixels. Because we use SvsS modality

![Fig. 2. Example images from different datasets: (a) VIPeR. (b) i-LIDS. (c) CAVIA4REID.](image-url)

Table 1. Results on VIPeR dataset.

<table>
<thead>
<tr>
<th>Top-k rank</th>
<th>VIPeR ((p = 316))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k = 1)</td>
<td>95.9</td>
</tr>
<tr>
<td>(k = 5)</td>
<td>94.30</td>
</tr>
<tr>
<td>(k = 10)</td>
<td>90.4</td>
</tr>
<tr>
<td>(k = 20)</td>
<td>87.5</td>
</tr>
<tr>
<td>(k = 30)</td>
<td>82.6</td>
</tr>
<tr>
<td>(k = 50)</td>
<td>78.48</td>
</tr>
</tbody>
</table>

![Table 1](table-url)
3.3. CAVIAR4REID

The CAVIAR4REID dataset contains person images in a shopping center from two different camera views. It contains 1221 images of 72 different pedestrians, out of which only 50 appear in both cameras and 22 persons come from the same camera. The number of each person’s images varies from 2 to 5 and the image size varies from 17*39 to 72*144. We chose the first and second image of each person and resize them to equal size. So in this dataset we have 144 images of 72 persons. The number of person used for test is then 36. Some examples are shown in Fig.2(c).

In this experiment m are set to be 9 and the dimension of WHOS n are set to be 9. The final feature dimension is then 18 and we use PCA to reduce it to be 9. The top-1 rank accuracy of our fusion strategy is 69.44%. The results are also similar: our three methods outperform the others although the improvement of fusion strategy is not very obvious. The results are shown in table 3 and Fig.3(c).

4. CONCLUSION

In this paper, we have developed a fusion representation for person re-identification and use KISSME algorithm for person matching. We combine mid-level color name descriptor and low-level WHOS descriptor in a fusion strategy. Experiments have demonstrated that our approach can significantly improve the recognition accuracy and achieve state-of-the-art results on several public person re-identification datasets. One thing to mention is that the number of person used for testing in i-LIDS and CAVIAR4REID is 59 and 36 respectively which is actually small when compared with VIPeR so the

![Fig. 3. Comparative performance evaluation on different datasets: (a) VIPeR. (b) i-LIDS. (c) CAVIAR4REID.](image-url)
improvement is not so obvious as in the latter case. We will go further to test our approach in more scalable datasets.

5. REFERENCES