Person Re-Identification by Spatial Pyramid Color Representation and Local Region Matching

Chunxiao LIU†(a), Nonmember, Guijin WANG(b), Member, Xinggang LIN(c), and Liang LI††(d), Nonmembers

SUMMARY Person re-identification is challenging due to illumination changes and viewpoint variations in the multi-camera environment. In this paper, we propose a novel spatial pyramid color representation (SPCR) and a local region matching scheme, to explore person appearance for re-identification. SPCR effectively integrates color layout into histogram, forming an informative global feature. Local region matching utilizes region statistics, which is described by covariance feature, to find appearance correspondence locally. Our approach shows robustness to illumination changes and slight viewpoint variations. Experiments on a public dataset demonstrate the performance superiority of our proposal over state-of-the-art methods.

key words: person re-identification, spatial pyramid, color histogram, covariance descriptor

1. Introduction

Person re-identification refers to matching images of the same person captured at different locations or across cameras. It has aroused much interest in recent years owing to its great significance in surveillance across camera networks, such as human retrieval and global understanding of person behavior. The surveillance scenarios envisaged raise severe challenges including illumination changes, viewpoint variations and occlusions, which result in appearance inconstancy of the same person in two images.

Current solutions can be categorized into two kinds. The first kind is to optimize the ranking of a true match, generally for a given dataset. Ensemble RankSVM [1] and probabilistic relative distance comparison [2] were developed to address the scalability and over-fitting problems caused by insufficient training samples. However, this kind of method is not extensible in a camera network because the scene structure and viewpoint may vary greatly for each camera. Thus training a model suitable for all camera pairs is not practical, if the feature used isn’t robust to viewpoint changes. The second kind is to seek a distinctive signature of human appearance, by means of exploiting possible features such as color [3], [4], texture [5], shape [6], local features [7], etc. In [3], symmetry-driven weighted color histogram and maximally stable color regions were proposed as color representations. Meanwhile a local feature describing highly recurrent texture characteristics was developed to complement the color feature. Spatial relation of features haven been explored recently. Occurrences of color [4] and shape [6] across human appearance were investigated to model the spatial relations so as to maintain robustness to appearance deformation. In [8], they extract region descriptors from a collection of grids. Each grid segment the object into a different number of subrectangles of equal sizes. Then similarity between two objects is computed by averaging the distances of the most similar rectangles in the grids. In [7], spatial pyramid matching kernel is employed to organize similarities between parts in different scales. Covariance is used to describe parts. However, in both [8] and [7], their features are extracted in local grids (or regions) separately without combined.

In this paper, we focus on appearance-based person re-identification. Our contributions are two fold. Firstly, a novel spatial pyramid color representation (SPCR) is proposed, to improve the discriminability of the color histogram by incorporating spatial information. Secondly, a local region matching scheme using covariance descriptor is proposed. It is robust to illumination changes and slight viewpoint variations. SPCR and local covariance depict the global and local properties of human appearance respectively. By combining these two complementary features, our method outperforms state-of-the-art methods on a public dataset.

The remainder of this paper is organized as follows. In Sect. 2, we introduce our spatial pyramid color representation and local region matching scheme, followed by their distance measures and combination for person re-identification. In Sect. 3, our proposed method is evaluated on a public dataset and compared with state-of-the-art methods. Section 4 gives the conclusion.

2. The Approach

In this paper, our main concern is to generate a distinguishing human signature for matching. Firstly, we roughly partition human body into three parts (head, torso and leg) vertically with a fixed ratio. Then distance between two persons is given by a weighted sum of distance scores of corresponding parts. For each part, we extract a color feature describing the whole part and a set of covariance features describing local regions. Thus distance between two parts is derived based on two measurements given by color feature
comparison and local region matching respectively.

2.1 Spatial Pyramid Color Representation

Color layout is an important and effective clue for the representation of human appearance, since sole color histogram cannot differentiate two different patterns with the same statistical color distribution. Therefore to augment color histogram with pixels’ locations, we partition the image spatially into subregions in a coarse-to-fine manner using spatial pyramid [9], and aggregate the color histograms over these subregions to form a holistic representation. We term our proposal spatial pyramid color representation (SPCR).

The SPCR is calculated as follows: let \( I \) denote an image. Color values of \( I \) are quantized into \( N \) centers, consequently an image is converted to a label map whose values are center indexes of the corresponding color values. Consider that we make a pyramid partition of label map into \( L \) levels. In level \( l \) of the pyramid, the grid has \( 2^l \) cells in both horizontal and vertical directions, as shown in Fig. 1. We denote the color histogram of region \( r \) in level \( l \) as \( H_l^r \), where \( 1 \leq r \leq 4^l \). To emphasize finer spatial information in higher pyramid level, the values of histograms in level \( l \) are multiplied by a weight which is set to \( \frac{1}{2^l} \). By concatenating the histogram sequences across the pyramid, we define our SPCR for image \( I \) as:

\[
SPCR(I) = [I^0H^0, I^1H_1^1, I^1H_2^1, \ldots, I^LH_n^L]
\] (1)

Compared to similar works [8] and [7], our usage of spatial relation is different in two aspects: on one hand, we incorporate spatial relation in the feature level to form a new holistic representation; on the other hand, features are weighted according to their spatial scales in our approach while they are treated equally in [8].

2.2 Local Region Matching

Within the context of person re-identification, S.Bak et al. [7] have firstly utilized covariance to describe human body parts which are segmented by a pre-trained part detector. However, viewpoint and pose changes in a camera network can affect the performance of a part detector. From a different perspective, we explore local regions for matching. We also use covariance descriptor to encode local region. The local region matching scheme consists of three steps: local region selection, descriptor generation and matching.

2.2.1 Region Selection

We perform a random sampling for region selection. Specifically, for each part, we first sample enough region candidates uniformly by assigning their centers and sizes randomly. Then these candidates are weighted according to their distances to the symmetry axis of human body in vertical direction. Regions with larger weights are selected considering that they are more robust to viewpoint variations.

In our implementation, we first sample a total of 100 regions for all the parts, and keep the top 50 regions sorted by weight in descending order.

2.2.2 Covariance Descriptor

For a given region \( R \subset I \), let \( \{z_m\}_{m=1}^{M} \) be the feature vectors extracted from \( M \) pixels inside \( R \). The covariance descriptor of region \( R \) is derived by:

\[
C_R = \frac{1}{M-1} \sum_{m=1}^{M} (z_m - \mu)(z_m - \mu)^T
\] (2)

where \( \mu \) denotes the mean vector of \( \{z_m\} \). In this paper, we use the following features to reflect information of each pixel:

\[
z = [H, S, V, I_x, I_y, I_{xx}, I_{yy}]
\] (3)

where \( H, S, V \) are the HSV color values. The image derivatives are calculated through the filters \([-1, 0, 1]^T\) and \([-1, 2, -1]^T\). Thus the covariance descriptor is a \(7 \times 7\) matrix.

2.2.3 Covariance based Region Distance

First we define the distance measure between two regions \( R_1 \) and \( R_2 \):

\[
d(R_1, R_2) = d(C_1, C_2) \times d(G_1, G_2)
\] (4)

where \( C_i \) and \( G_i \) are the covariance descriptor and geometry information of \( R_i \). We use the distance measure introduced in [10] for two covariance matrices:

\[
d(C_1, C_2) = \sqrt{\sum_{i=1}^{T} \ln^2 \lambda_i(C_1, C_2)}
\] (5)

where \( \lambda_i(C_1, C_2) \) are the generalized eigenvalues of \( C_1 \) and \( C_2 \), computed from

\[
\lambda_i C_1 x_i - C_2 x_i = 0 \quad (t = 1 \ldots T)
\] (6)

and \( x_t \neq 0 \) are the generalized eigenvectors. To impose geometry constraints when comparing two regions, we define

\[
d(G_1, G_2) = 1 - e^{-\|G_1 - G_2\|_2}
\] (7)

where \( G_i = [x_i, y_i, W_i, H_i]^T \) represents the centroid and scale of region \( R_i \).
2.2.4 Local Region Matching

Local region matching is performed for part comparison. In each body part, multiple regions are sampled and a set of covariance descriptors is generated. Let $I_{np}^{sp}$ denote the $np$-th part of person $i$, $R(I_{np}^{sp})$ denote a set of local regions sampled from $I_{np}^{sp}$, and $d(R(I_{np}^{sp}), R(I_{np}^{sp}))$ denote the distance between the $np$-th part of two persons. Specifically, the $d(R(I_{np}^{sp}), R(I_{np}^{sp}))$ is represented by the sum of $K$ minimum distances between region set $R(I_{np}^{sp})$ and $R(I_{np}^{sp})$.

2.3 Combined Features for Person Re-identification

Dissimilarity of two persons is estimated by combining the color feature distance and local region matching:

$$D(I_1, I_2) = \omega \cdot d(SPCR(I_1), SPCR(I_2)) + (1 - \omega) \cdot d(R(I_1), R(I_2))$$

(8)

where $\omega$ denotes the weight, $\omega \in (0, 1)$. Bhattacharyya distance is employed to measure the dissimilarity on SPCR. We combine the feature scores of all the parts:

$$d(SPCR(I_1), SPCR(I_2)) = \sum_{np} w_{np} d(SPCR(I_1^{np}), SPCR(I_2^{np}))$$

(9)

$$d(R(I_1), R(I_2)) = \sum_{np} w_{np} d(R(I_1^{np}), R(I_2^{np}))$$

(10)

where $w_{np}$ denotes the weight of the $np$-th part.

3. Experiments

We test our algorithm on the ETHZ dataset [5] that is commonly used for person re-identification. For each person, a set of images varied from a few to hundreds, is extracted from videos. There are 3 sequences: Seq.1 has 83 persons, for a total of 4857 images; Seq.2 has 35 persons, for a total of 1936 images; Seq.3 has 28 persons, for a total of 1762 images. The challenges of this dataset lie in the illumination changes, occlusions, scale changes, and slight viewpoint variations. We make comparison with other three methods, denoted for short by SDALF [3], SCR [7] and CG [8], respectively. For SDALF, we use the result reported in [3] which showed best performance on this dataset. We implement SCR and CG strictly following the details stated in their papers, and use their default parameters. Cumulative Matching Curve (CMC) is used for evaluation, which represents the correct matching rate in the top $N$ ranks.

All the parameters are set as follows: images are normalized to $64 \times 32$ pixels. A fixed part partition (head/torso/leg) ratio is set as $[0.2, 0.3, 0.5]$ in vertical dimension, with their weights fixed as $[0, 0.7, 0.3]$ as we neglect the information on head. SPCR is extracted in HSV color space. The numbers of bins for HSV channels are set to $[16, 16, 4]$ respectively. As for feature combination, $\omega$ is set to 0.8. We use the foreground mask provided by [3] and extract feature only on human body. As to SCR and CG, spatial scale is set to 2 when both of these methods show best results.

In this experiments, given a gallery set $A$ and a probe set $B$, we are to associate the person in $B$ to the corresponding person in $A$. We randomly pick an image for each person to form the gallery set, and the rest are used as probe set. This procedure is repeated 10 times as that in [3] and the mean CMCs are reported. Figure 2 shows the result of SDALF and ours (denoted by ‘SPCR (L=1)+LRM’). We achieve performance gain of about 5% in rank 1 of Seq.1 and Seq.3. Figure 2 also illustrates that recognition rates have been effectively raised when SPCR is used. SPCR (L=0) means no spatial pyramid is built. Specifically, about 10% has been prompted in rank 1 recognition rate when one layer (L=1) is added. But the finer spatial partition doesn’t promise better result, as denoted by the performance degradation of SPCR (L=2) in Seq.3. The choice of pyramid layer depends on image resolution, and the extent of viewpoint or illumination variations. When the two proposed features are combined, the overall performance is boosted, which shows these two features can be well integrated. Comparison with similar works is presented in Fig. 3. Compared to CG using a same scale, our SPCR (L=2) has an increase of %7 and %15 in rank 1 of Seq.1 and Seq.2 respectively. However, CG provides better results in Seq.3 where viewpoint...
change is a little severe. Our local matching scheme also produces better recognition rates than SCR in all sequences. One possible reason is that SCR assumes a correct part correspondence can always be found (e.g. left arm to left arm). This doesn’t hold true when occlusion or viewpoint variation happens. In contrast, our LRM uses reliable region correspondences by searching all the candidate regions, which is more reasonable to solve these difficulties.

4. Conclusion

In this paper, we address the person re-identification problem by providing a novel distinctive appearance representation. It consists of two complementary elements: the Spatial Pyramid Color Representation which effectively incorporates spatial information into color histogram, and the Local Region Matching that employs covariance of local regions. This method is robust to illumination changes, scale changes and slight viewpoint variations. However, drastic viewpoint and pose variations aren’t well handled, which would be the main focus of our future work.

Acknowledgments

This research project was supported by Sony China Research Laboratory, Sony (China) Ltd. and NSFC 61132007.

References