J. Vis. Commun. Image R. 29 (2015) 89–102

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

A person re-identification algorithm by exploiting region-based feature salience ☆,☆☆

Yanbing Geng [a,c], Hai-Miao Hu [a,b,*], Guodong Zeng [a], Jin Zheng [a,b]

[a] Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[b] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China
[c] College of Electronic and Computer Science and Technology, North University of China, Taiyuan, Shanxi 030051, China

ARTICLE INFO

Article history:
Received 29 August 2014
Accepted 4 February 2015
Available online 19 February 2015

Keywords:
Person re-identification
Region-based feature saliance
Salient color descriptor
Feature extraction
Feature fusion
Metric distance calculation
Illumination variation
Video surveillance

ABSTRACT

Due to the changes of the pose and illumination, the appearances of the person captured in surveillance may have obvious variation. Different parts of persons will possess different characteristics. Applying the same feature extraction and description to all parts without differentiating their characteristics will result in poor re-identification performances. Therefore, a person re-identification algorithm is proposed to fully exploit region-based feature salience. Firstly, each person is divided into the upper part and the lower part. Correspondingly, a part-based feature extraction algorithm is proposed to adopt different features for different parts. Moreover, the features of every part are separately represented to retain their salience. Secondly, in order to accurately represent the color feature, the salient color descriptor is proposed by considering the color diversity between current region and its surrounding regions. The experimental results demonstrate that the proposed algorithm can improve the accuracy of person re-identification compared with the state-of-the-art algorithms.

1. Introduction

Person re-identification is the task of establishing correspondences between observations of the same person in different videos. It faces many challenging issues like low resolution frames, time-varying light conditions, pose changes and partial occlusions under the uncontrolled complex environment. Conventional biometric traits such as face and gait are infeasible or unreliable owing to the scarce resolution of sensors. Usually, it is assumed that people among the different camera views can be effectively identified by some apparent information such as color, texture, edge and structure. The appearance-based person re-identification is widely used in the state-of-the-art algorithms.

Appearance-based person re-identification algorithm can be divided into two groups, namely the metric learning algorithm [20–29] and the feature representation algorithm [1–19,41–42]. Aiming to map the original feature to a new feature space, the first group learns a proper distance metric, which makes sure the feature distance between different instances for the same person is small in the new feature space. An optimal distance metric is learned through maximizing the inter-class variation while minimizing the intra-class variation in the literature [20,23–25], but it could easily cause over-fitted under the limited training samples. Zheng et al. [29] maximize the probability of a pair of true match having a smaller distance than that of a wrong match pair, which can alleviate the over-fitting problem. Xiong et al. [21] design classifiers to learn specialized metrics, which enforce features from the same individual to be closer than features from different individuals. However, enormous labeled training images should be obtained for the above mentioned methods, which may be hardly to implement in practice. Aiming at this deficiency, Liu et al. [22] propose a semi-supervised coupled dictionary learning, both labeled and unlabeled images are jointly learned in the training phase. However, these metric learning algorithms are inapplicable for practical surveillance applications, since the retraining must be carried out for per-dataset or per camera-pair.

The feature representation algorithms focus on seeking a distinctive and stable feature expression. Typical visual features are extracted for person re-identification, such as color, texture and shape. These features are always combined to improve the recognition rate. Feature extraction and multi-feature fusion are two

☆ This paper has been recommended for acceptance by M.T. Sun.

☆☆ This work was partially supported by the National Science Fund for Distinguished Young Scholars (No. 61125206), the National Natural Science Foundation of China (No. 61370121), the National Hi-Tech Research and Development Program (863 Program) of China (No. 2014AA015102), and Outstanding Tutors for doctoral dissertations of S&T project in Beijing (No. 20131000602).

* Corresponding author at: Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China.
E-mail address: frank0139@163.com (H.-M. Hu).

http://dx.doi.org/10.1016/j.jvcir.2015.02.001
1047-3203/

main issues for the feature representation. We will analyze them in details.

Visual feature can be roughly categorized into global and local features. Color is the most commonly appearance feature for person re-identification. Global color features are encoded via chromatic histogram, since global color features often fail to characterize color distributions within the body part. Local color features are used to localize the color information through the patches segmentation. Mean-shift algorithm is used to segment an image into a series of patches in the literature [2], each patch has similar colors and HSV components are combined to represent the color information. However, this algorithm ignores the structure information of the person body, which will lead to the mismatch between torsos and legs. In order to deal with this problem, structure information is utilized as the spatial constraint by dividing a body into several asymmetrical parts (e.g., head, torso and legs) in the literatures [1,8–10]. According to [5], different parts are implicitly isolated and patches located in different parts are matched individually, which can efficiently reduce the mismatch between different parts. Asymmetry-based Histogram Plus Epitome (AHPE) feature is proposed to represent the global color histogram and local epitome information against the low resolution, occlusion and illumination variation in the literature [9]. However, the above algorithms often fail to differentiate persons with similar clothes and trousers. Inspired by human eyes of pedestrian identification relying on some salient regions, Zhao et al. [6] identify the matching pairs by means of the image salient patches. In their work, the salient patches are detected in an unsupervised manner and incorporated in patch matching to find reliable matches. But the salient patches may change due to illumination variation. Yang et al. [4] propose a salient color name based color descriptor (SCNCD) to analyze images by the semantic information, based on SCNCD, color distributions over sixteen color names in different color spaces are fused to address the illumination problem. Furthermore, because color is easily influenced by illumination variations across camera views, texture and shape are usually combined to model the human appearance. Schwartz and Davis [11] combine color, texture and edge features to represent the appearance, the partial least squares (PLS) is used for dimension reduction. But it fails to differentiate the salience among different features during the dimension reduction process.

Generally, the existing feature representation algorithms are insufficient in the following two aspects. Firstly, for the feature extraction, same features are adopted to describe different parts, on the assumption that these features are optimal for all parts. Since the person appearances captured by different cameras have obvious variations of posture and viewpoint, different features can be adopted for each part according to their respective characteristics. Furthermore, during the features fusion, there are various intrinsic meanings for each fused feature. Current multi-feature fusion algorithms fail to differentiate the salience among different features, which cannot make full use of the contribution of each feature.

Secondly, two kinds of local color descriptor, the stable patch color descriptor and the salient patch color descriptor are always used to describe the apparent color. However, both of them are less discriminative between the different appearances in the case of similar dress [1,6]. The salience should be considered not only for the local patch but also for the local color information. The detailed analysis of these two observations will be further discussed in Section 2.

Therefore, this paper proposes a person re-identification algorithm by fully exploiting the region-based feature salience (labeled as "RbFS"). In order to use the structural and spatial information, one human body is divided into the upper part and the lower part, and each part is further divided into multiple patches. The contributions of the proposed algorithm are two aspects. On the one hand, a part-based feature extraction algorithm is proposed to adopt different features for different parts according to the feature effectiveness and the part characteristics, so that the features for different parts are separately represented and each feature can retain its intrinsic meanings and salience. On the other hand, the salient color descriptor is proposed to embody the color representation and discrimination by detecting the salient color patch and considering the color diversity between current patch and its surrounding patches. The proposed algorithm is extensively evaluated on several public datasets, wherein images are captured in the video surveillance. The experimental results demonstrate that the proposed algorithm can significantly improves the accuracy of person re-identification compared with the-state-of-the-art algorithms.

The remainder of this paper is organized as follows. Two observations are elaborated in Section 2. The proposed algorithm is described in detail in Section 3. Section 4 summarizes the proposed algorithm. Section 5 evaluates the proposed algorithm on four datasets, including CAVIAR4REID [33], VIPeR [19], i-LIDS [39] and ETHZ [32]. And the paper is concluded in the final section.

2. Observations and justifications

In this section, two observations are discussed in details and some experiments are carried out for justifications.

2.1. Feature extraction and fusion

The assumption of features universally optimal for the whole body is not desirable owing to the different characteristics of torso and legs. It is obvious from Fig. 1 that the lower region (e.g., legs) varies seriously whereas the upper region (e.g., torso) maintains relatively stable in general walking behavior.

Commonly, the appearance of human is usually characterized in three aspects, color, shape and texture. Color has proven to be effective for the task of person re-identification. It remains stable to the variations of posture and viewpoint even at lower resolutions. Texture and structure-shape information are complemented when color information degrades under illumination change. Each feature has its discriminative power, it is not powerful enough to characterize all apparent parts. Therefore the feature salience and effectiveness should be considered to represent different apparent part during the feature extraction. It is necessary for different regions to adopt different features. This conclusion can be proved through the following experiment on ETHZ and VIPeR datasets. For ETHZ, we randomly sampled 12 persons from each sequel. Six images for each person are selected, one for gallery set and the remainder for probe set. In addition, 100 persons are randomly chosen from VIPeR. For each person, one image forms gallery set and the other forms probe set.

In this experiment (as shown in Fig. 1), we apply human 2D vertical ellipse models [36] to partly remove the influence of the different background clutters, and then each body is divided into the upper part and the lower part by using adaptive body segmentation [5]. In order to make full use of the local detail information, each part is further segmented into several patches with Mean-shift [2]. Color feature (e.g., HSV value), texture features (e.g., Uniform Pattern Local Binary Patterns histogram, UPLBP [40]) and oriented gradient features (e.g., Histograms of Oriented Gradients, HOG [31]) are selected to describe each patch. Table 1 presents the rank 1 matching rate by using different feature on different parts. For the upper part, the average recognition ratios of adopting HSV value and HOG descriptor are 76% and 58% respectively, they are higher than those of adopting UPLBP. For the lower region, the