Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Fusion of multiple channel features for person re-identification

Xuekuan Wang [a,1], Cairong Zhao [a,b,c,1], Duoqian Miao [a,*], Zhihua Wei [a], Renxian Zhang [a], Tingfei Ye [c]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[b] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information (NJUST), Ministry of Education, Nanjing 210094, China
[c] Key Laboratory of Cloud Computing and Intelligent Information Processing of Changzhou City, Jiangsu University of Technology, 213001 Changzhou, China

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form
30 November 2015
Accepted 13 December 2015

Keywords:
Person re-identification
Multiple channel
Feature fusion

ABSTRACT

Person re-identification plays an important role for automatic search of a person's presence in a surveillance video, and feature representation is a critical and fundamental problem for person re-identification. Besides, an reliable feature representation should effectively adapt to the changes of illumination, pose, viewpoint, etc. In this paper, we propose an effective feature representation called fusion of multiple channel features (FMCF) which captures different low-level features from multiple channels of HSV color space, considering the characteristics of different color channels and fusing color, texture and correlation of spatial structure. Furthermore, it takes advantage of an overlapping strategy to eliminate contrast of local cells in an image. In addition, we apply the simple weight distance metric to measure the similarity of different images, rather than metric learning which relies on a specific feature and requires more computing resources. Finally, we apply the proposed method of FMCF on the i-LIDS Multiple-Camera Tracking Scenario(MCTS) and CUHK-01person re-identification datasets, and the experimental results demonstrate that it is more robust to the variation of visual appearance.

1. Introduction

The task of person re-identification (re-id) is to match pedestrian images observed from disjoint views in non-overlapping camera networks [1]. It has attracted more and more attention in recent years due to its important applications in video surveillance. The key issue of re-id system is to capture reliable and robust features from pedestrian images and measure the similarity among them to estimate if they are from the same person. However, the complexity of the environment, which is affected by illumination, pose, viewpoint, occlusion, image resolution and camera setting in non-overlapping camera system, leads to various challenges [2]. At present, the state-of-the-art approaches for person re-identification are mainly divided into two groups: (1) the appearance-based approach designing of distinctive and stable descriptors to represent the person's appearance; (2) the metric learning approach obtaining a suitable metric method which minimizes the distance of the same person and maximizes the distance of different persons.

Most existing appearance-based approaches concern low-level features such as color (color histogram, Dominant color, color space, etc. [3–5,45]), texture (local binary pattern (LBP), Gabor, Co-occurrence matrix [6-8,35,40], etc. and shape [9–11,44,47]. These features are always combined to improve the recognition rate. Aiming to seek a distinctive and stable feature expression, researchers have proposed a lot of feature representation algorithms, ranging widely from symmetry-driven accumulation [12], covariance descriptor [13], horizontal stripe-based partition [14], pyramid matching [15], graph matching [16,48], salience matching [10,17], local maximal occurrence [18], hash model [41], sparse learning model [46,49,50], deep learning model [19–21], etc. Meanwhile, feature extraction and multi-feature fusion are two main issues for feature representation. In [22], D. Gray and H. Tao propose the method of ensemble of local features (EFL), achieving an efficient and intelligent descriptor for viewpoint invariant pedestrian recognition. While, in [23], the authors design the approach of bag of ensemble colors to combine low-level color histogram and semantic color names to represent human appearances. These handcrafted or learning-based descriptors have made impressive improvements over robust feature representation, and advanced the person re-identification research. Unfortunately, it is also extremely difficult to extract a stable feature which effectively adapts to serve changes and misalignment across disjoint views.

Another aspect of person re-identification is how to learn a robust distance or similarity function to deal with the complex matching problem. Many metric methods simply choose a standard distance such as l_1–norm [24], l_2 – norm – based distance

* Corresponding author.
E-mail addresses: zhaocairong@tongji.edu.cn (C. Zhao), dqmiao@tongji.edu.cn (D. Miao).

[1] The authors contribute equally to this work.

http://dx.doi.org/10.1016/j.neucom.2015.12.140
0925-2312/

[25], or Bhattacharyya distance [22]. However, they would essentially treat all features equally without learning and discarding bad features selectively, thus the matching result is always undesirable. In contrast, the distance learning based approaches typically learn a discriminative metric between appearance features of the same person and different persons across camera pairs. These methods mainly include Rank SVM [26], least squares [43], Relative Distance Comparison (RDC) [14], Kernel-Based Metric [27], Mahalanobis distance learning [29], Deep Metric Learning (DML) [30], metric ensembles [28], non-convex model [42], and Iterative re-weight sparse ranking [31] etc. In practice, many previous metric learning methods show a two-stage procedure for metric learning, that is, the Principle Component Analysis (PCA) is first applied for dimension reduction, then metric learning is performed on the PCA subspace. However, this two-stage procedure may not be optimal for metric learning in a low-dimensional space, because samples with different classes may already be cluttered after the first stage.

In this paper, we propose an effective appearance-based feature representation called fusion of multiple histogram features (FMCF). Different from the previous works, FMCF pays more attention to the properties with different channels of color space *HSV* and captures the weight-color histogram, texture and spatial structural information using color space component instead of gray-value images. Main contributions in the current paper are as follows:

A. A new feature descriptor called fusion of multiple channel features (FMCF) has been proposed.
B. The proposed method of FMCF captures color information form hue and saturation components of HSV color space, while texture and spatial structural information are extracted from value component.
C. Joint histogram, acquisition and matching have been done for person re-identification.
D. Histogram is captured with overlapping strategy from three components of HSV color space.

The remainder of this paper is organized as follows. We review the related works in Section 2. We introduce the theory of the proposed approach in Section 3. In Section 4, we carry out the comparative experiments on two public person re-identification datasets and give the detailed discussion based on the experimental results. Finally, conclusions are offered in Section 5.

2. Related work

This paper aims to seek an effective appearance-based method for person re-identification from the view of multiple channel features extraction. Therefore, first of all, we present an overview of the relevant works, i.e., HSV color space, local binary pattern (LBP) [33] and histogram of Oriented Gradients (HOG) [34].

2.1. HSV color space

In general, there are three types of images, including binary image, gray scale image and color image. Aiming to reduce the complexity of algorithm, most approaches just concern the gray scale image. In contrast, the color image consists of multiple channels which contain a range of intensity and describe more rich information. Among them, the RGB color space composed of red, green and blue is widely utilized for image representation. However, the three channels pay more attention to the color property, ignoring other characteristics of color space. Hence, in our proposed approach, we consider the HSV color space that stands for hue, saturation and value which describe different characteristics of color image.

Hue component is directly related to the color which can be distinguished by the human eye. It is defined as an angle value, varying from 0° to 360°. Each number corresponds to a different color. Saturation component describes the purity of color component, and the value shows the intensity of a color. It is numbered from 0 to 1, as it goes from low to high intensity of color. Meanwhile, value component also varies from 0 to 1 and it is most similar to gray-scale image.

Compared with other color spaces, many research works have explained that HSV color space is more effective for extracting color, intensity and brightness from images [38]. Besides, different channels of HSV color space are relatively independent and have weak correlation, which is conducive to extracting more varied and sufficient information from single color channel. Meanwhile, it can ensure that the dimension of feature vector is low. In our proposed approach, images are converted from the RGB space to the HSV color space, besides we deal with the characteristics of different channels to capture suitable and abundant features which combine the color, texture and spatial structural information for person re-identification.

2.2. Local binary pattern (LBP)

A landmark representative of these structural image descriptors is Local Binary Pattern first proposed by Ojala et al. [33] as a gray-scale invariant texture descriptor. LBP code is obtained by thresholding its circularly symmetric n-neighbors in a circle of radius r with a pixel value of central point, and arranging the results as a binary string, shown in Fig. 1. It is stable and robust for the change of illumination. The mathematical representation of *LBP* is as follows:

$$LBP_{n,r} = \sum_{i=1}^{n} P_1(I_i - I_c) \times 2^{i-1} \tag{1}$$

$$P_1(t) = \begin{cases} 1, & t \geq 1 \\ 0, & else \end{cases} \tag{2}$$

$LBP_{n,r}$ obtains the local binary pattern of each pixel, where n and r are the number of neighboring pixels and the radius of circle, taken for computation. I_i is the pixel value of center point and I_c is the pixel value of neighboring pixels.

2.3. Histogram of Oriented Gradients (HOG)

The Histogram of Oriented Gradients (HOG) is a popular descriptor that was initially proposed for pedestrian detection by Dalal and Triggs [34]. HOG is represented by the 3D histogram of gradient locations and orientations, and employs both rectangular and log-polar location grids. The process of generating HOG Descriptor for an image is shown in Fig. 2. Aiming to capture the HOG feature, the gradient and director of the images are compute by Eqs. (3)–(6).

$$G_x(x, y) = I(x+1, y) - I(x-1, y) \tag{3}$$

$$G_y(x, y) = I(x, y+1) - I(x, y-1) \tag{4}$$

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \tag{5}$$

$$\alpha(x, y) = \tan^{-1}(G_x(x, y)/G_y(x, y)) \tag{6}$$

There, $I(x, y)$, $G_x(x, y)$, $G_y(x, y)$, $G(x, y)$ and $\alpha(x, y)$ represent color pixel value, the gradient of horizontal director, the gradient of vertical director, gradient values and directions at point (x, y),