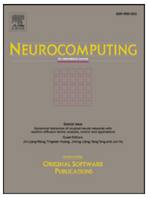
Neurocomputing 000 (2018) 1–13

Contents lists available at ScienceDirect

Neurocomputing

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



Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning

Husheng Dong[a,b], Ping Lu[b], Shan Zhong[c], Chunping Liu[a,d,e], Yi Ji[a], Shengrong Gong[c,a,f,*]

[a] School of Computer Science and Technology, Soochow University, Suzhou, China
[b] Suzhou Institute of Trade and Commerce, Suzhou, China
[c] Changshu Institute of Science and Technology, Changshu, China
[d] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China
[e] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, China
[f] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

ARTICLE INFO

Article history:
Received 29 October 2016
Revised 26 February 2018
Accepted 8 April 2018
Available online xxx

Communicated by Dr Jiwen Lu

Keywords:
Person re-identification
Feature representation
Metric learning
Generalized similarity

ABSTRACT

To solve the challenging person re-identification problem, great efforts have been devoted to feature representation and metric learning. However, existing feature extractors are either stripe-based or dense-block-based, the fine details and coarse appearance are not well integrated. What is more, the metrics are generally learned independently from distance view or bilinear similarity view. Few works have exploited the mutual complementary effects of their combination. To address these issues, we propose a new feature representation termed enhanced Local Maximal Occurrence (eLOMO) which fuses a new overlapping-stripe-based descriptor with the Local Maximal Occurrence (LOMO) extracted from dense blocks. Such integration makes eLOMO resemble the coarse-to-fine recognition mechanism of human vision system, thus it can provide a more discriminative descriptor for re-identification. Besides, we show the advantages of learning generalized similarity by combining the Mahalanobis distance and bilinear similarity together. Specifically, we derive a logistic metric learning method to jointly learn a distance metric and a bilinear similarity metric, which exploits both the distance and angle information from training data. Taking advantage of learning in the intra-class subspace, the proposed method can be solved efficiently by coordinate descent optimization. Experiments on four challenging datasets including VIPeR, PRID450S, QMUL GRID, and CUHK01, show that the proposed method outperforms the state-of-the-art approaches significantly.

1. Introduction

Person re-identification is the task of matching individuals across disjoint camera views over distributed spaces, which plays an important role in intelligent video surveillance. Although it is assumed that people do not change clothes in different camera views, person re-identification still remains a challenging problem due to large appearance variations caused by illumination, pose, viewpoint, and occlusion.

Great efforts have been devoted for years to tackle person re-identification along two directions. One is to design robust visual descriptors against cross-view variations, and the other is to learn a discriminant similarity/distance functissson to determine whether an image pair belongs to the same person or not. For visual descriptors, a number of feature representations have been proposed, such as Symmetry-Driven Accumulation of Local Features (SDALF) [1], Mid-level Filter (MLF) [2], Biologically Inspired Features (BIF) [3], Salient Color Names (SCN) [4], Local Maximal Occurrence (LOMO) [5], and the Gaussian of Gaussian (GOG) descriptor [6]. Most of them are extracted from either horizontal stripes or dense blocks. Although impressive advancement has been made, designing a more robust yet discriminative descriptor remains an open problem.

As for similarity/distance function learning, a number of metric learning algorithms have been devised [5,7–14]. Some of them, like [10,11,13,14], focus on learning a Mahalanobis distance metric from distance constraints. While some other works, like [7,12], seek for a bilinear similarity metric by utilizing the angle information between instances in high-dimensional feature space. However, most

[*] Corresponding author at: Changshu Institute of Science and Technology, Changshu, China..
E-mail address: shrgong@cslg.edu.cn (S. Gong).

https://doi.org/10.1016/j.neucom.2018.04.013
0925-2312/© 2018 Published by Elsevier B.V.

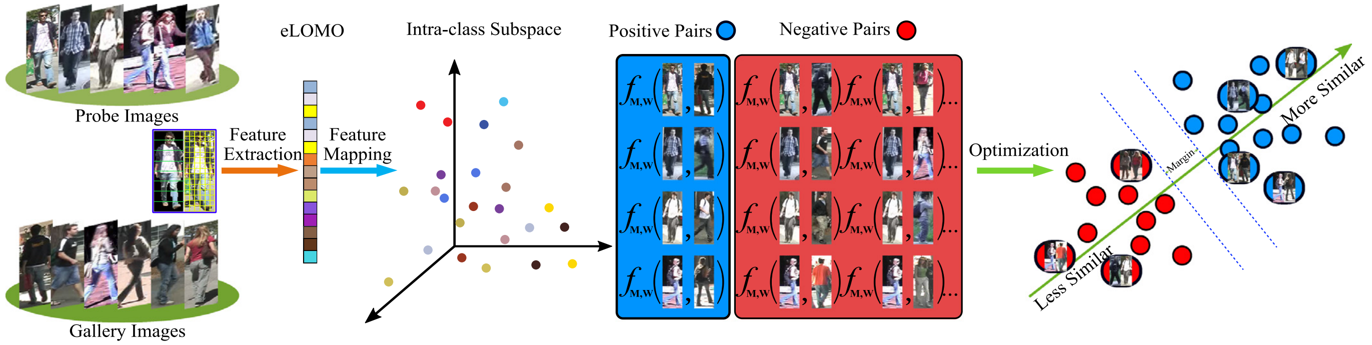


Fig. 1. The pipeline of the proposed method. The eLOMO features are extracted from every image first, and then they are mapped into the intra-class subspace. The generalized similarity function $f_{\mathbf{M},\mathbf{W}}$ is trained by maximizing intra-class similarities and minimizing inter-class similarities.

of the works fail to exploit the mutual complementary effects of their combination. Only considering either of them may lead to a less discriminative similarity measurement.

In this paper, we propose an efficient feature representation termed enhanced Local Maximal Occurrence (eLOMO), and a Generalized Similarity Metric Learning (GSML) method for person re-identification. The eLOMO is an integration of a new overlapping-stripe-based descriptor with the existing LOMO [5] feature. The stripe-based descriptor can better exploit coarse appearance information from larger regions, while LOMO is good at capturing the fine details of dense blocks. Thus the fusion of them can lead to a coarse-to-fine representation which is in line with the human recognition mechanism. To learn a discriminant similarity function, we combine the Mahalanobis distance and the bilinear similarity together, such that the distance and angle information of training data are exploited simultaneously. The proposed method is formulated as a logistic metric learning problem with Positive Semi-definite (PSD) constraints, and we derive an efficient coordinate descent algorithm to solve it based on the Accelerated Proximal Gradient (APG) optimization method. To suppress the large intra-class variations of cross-view appearances, we project samples into the intra-class subspace before learning. The pipeline of the proposed method is shown in Fig. 1.

We conduct extensive experiments to validate the efficacy of the proposed method. Experimental results show that the proposed method achieves significant improvements over existing approaches on four challenging person re-identification datasets, namely VIPeR [15], PRID450S [16], QMUL GRID [17], and CUHK01 [2].

The rest of this paper is organized as follows. In Section 2 we briefly review the related works and discuss their differences with our method. Section 3 introduces the eLOMO feature representation. Section 4 presents the GSML in detail. The experimental results and the analysis of our method are reported in Section 5. Finally, we draw some conclusions in Section 6.

2. Related work

Given one probe image containing an individual of interest, the task of person re-identification is to find its true match (or usually the best match) from a large number of gallery images. Existing works for solving this problem generally follow a two-step paradigm. Firstly, a robust and distinctive feature representation is extracted for every pedestrian image. Secondly, the similarity/distance for each probe-gallery image pair is measured by a certain metric, which is then used to rank the gallery images for each probe. Majority of existing methods focus on either designing feature representations or learning discriminative metrics. Here we only briefly review some related works, more comprehensive surveys can be found in [18–21].

2.1. *Feature representations for person re-identification*

Many approaches try to build distinctive feature representation for describing pedestrian appearance in different environments. In order to achieve discrimination and robustness against different variations, they are generally extracted from horizontal stripes [4,15,22,23] or dense blocks [5,8,9,24,25]. For example, the Ensemble of Local Features (ELF) [15] and the SCN [4] are extracted from six non-overlapping horizontal stripes. As an extension of ELF, the ELF18 [22] is computed from 18 non-overlapping stripes. From overlapping stripes, Lisanti et al. [23] computed the weighted local features of color histogram, Local Binary Patterns (LBP), and Histogram of Gradient (HOG). In general, these stripe-based descriptors are robust to the cross-view body misalignment problem and can well capture the holistic appearance information.

Compared to stripe-based descriptors, the features computed from dense blocks can better capture fine details from relatively small patches. By computing Gabor filters and covariance from dense grids, Ma et al. [3] proposed the BIF feature. Zhao et al. [2] tried to learn Mid-level filters from the clusters of dense patches. Recently, Liao et al. [5] proposed the LOMO descriptor which has shown impressive robustness against viewpoint changes. By describing each pedestrian image as a set of hierarchical Gaussian distributions represented by the means and covariance, Matsukawa et al. [6] designed the GOG descriptor which is also computed from dense blocks in essence. However, just as one coin has two sides, a disadvantage of these dense-block-based descriptors is that they are not good at describing the holistic appearance, although some of them have considered computing features from multi-scale spaces (e.g. LOMO and GOG).

To combine the advantages of both stripe-based and dense-block-based features, we argue that feature representations should be computed from stripes (larger regions) and dense blocks (small patches) simultaneously. In this work, we fuse the successful LOMO with the features extracted from a pyramid space of two-level overlapping stripes. As a consequence, the fine details from dense blocks and coarse appearance from larger regions are well integrated to boost the discrimination.

With the blossom of deep learning, there are also some works try to learn features by the powerful deep models, such as [26–28]. Even with the generic metric of $L2$ norm, impressive performance can be achieved. However, they require very large number of training data and can easily suffer the over-fitting

Please cite this article as: H. Dong et al., Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.04.013