BODY-STRUCTURE BASED FEATURE REPRESENTATION FOR PERSON RE-IDENTIFICATION

Hong Liu, Liqian Ma, Can Wang

Engineering Lab on Intelligent Perception for Internet of Things (ELIP),
Shenzhen Graduate School, Peking University, China
hongliu@pku.edu.cn, maliqian@sz.pku.edu.cn, canwang@pku.edu.cn

ABSTRACT

Person re-identification is valuable for intelligent video surveillance and has drawn wide attention. Although person re-identification research is making progress, it still faces some challenges such as varying poses, illumination and viewpoints. As a major aspect of person re-identification, feature representation has been widely researched. Low-level descriptors are generally used in existing works, which do not take full advantage of body structure information and result in low discrimination. In this paper, body-structure based mid-level feature representation is proposed, which introduces body structure pyramid for codebook learning and feature pooling. Additionally, low computational LLC is used to encode mid-level features. Experimental results on two challenging datasets VIPeR and CUHK01 have demonstrated that our approach outperforms the state-of-the-art methods.

Index Terms— Person re-identification, Feature representation, Body structure, Human appearance

1. INTRODUCTION

Person re-identification deals with the recognition of individual who appears in non-overlapping camera views, which is fundamental and essential for intelligent video surveillance. Generally, research difficulties lie in significant ambiguity brought by variations of poses, illumination and viewpoints, resulting in inaccuracy of person re-identification.

Recent years have witnessed lots of researches in this field and there are two major research aspects in person re-identification: feature representation and model learning [1]. Considering feature representation, high-level features such as gender and age are difficult to reliably acquire due to the unconstrained viewpoints of individuals as well as the impoverishment of visual information in real-world surveillance scenarios. Generally, low-level descriptors [2–4] are used to describe body appearance in most literatures, which commonly can not detailedly describe local information and results in limited discrimination. Person re-identification methods using mid-level features have shown better performance since mid-level features are insensitive to space misalignment and thus robust to variations of poses and viewpoints. Therefore, this paper focuses on mid-level feature representations.

Bag-of-features (BoF) is a classical mid-level feature representation framework which has demonstrated excellent performances in computer vision tasks such as image classification [5, 6] and action classification [7, 8]. However, the traditional spatial pyramid widely used in BoF [5, 6] does not consider body structure information which is significant for person re-identification. Actually, individuals in images have roughly consistent structures in vertical direction, $e.g.$ head in the top and legs in the bottom. This allows us to describe the structure information of individuals using a common approach. Besides, different body parts have different color and texture characteristics, which suggests corresponding representations for different body parts. Therefore, making full use of body structure information may improve the performance [2]. In this paper, a novel body-structure based feature representation (BSFR) approach for person re-identification is proposed. A new body-structure pyramid is put forward to represent body-structure information while Locality-constrained Linear Coding(LLC) [6], one extension of BoF, is utilized to encode low-level descriptors into mid-level representations.

Relation to prior work: Farenzena $et\ al.$ described body parts with low-level descriptors for re-identification, which results in limited discrimination [2]. Rui Zhao $et\ al.$ learned mid-level filters to represent features and achieved good performance [9]. However, this method doesn't make full use of body structure information and has a high cost of computation. Yang $et\ al.$ introduced LLC to encoded low-level descriptors into mid-level features, which has a low cost of computation and better discrimination but doesn't consider body structure information [10]. Taking all factors into account, our approach introduces body structure information for codebook learning and feature pooling, and uses mid-level features encoded by LLC.

This work is supported by National Natural Science Foundation of China(NSFC, No.61340046, 60875050, 60675025), National High Technology Research and Development Program of China(863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, CX-C201104210010A, JCYJ20130331144631730, JCYJ20130331144716089)

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

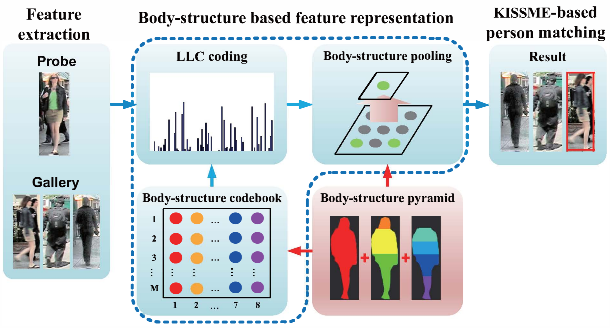


Fig. 1. The pipeline of our framework. The eight distinct colors in the body-structure pyramid are related to eight sub-codebooks in body-structure codebook. **Best viewed in color.**

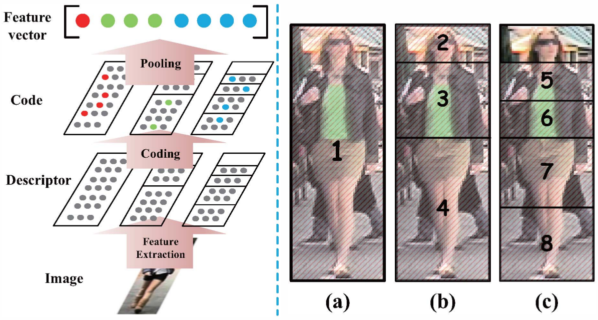


Fig. 2. Left: flowchart of the body-structure pyramid for pooling features. Right: our proposed body-structure pyramid composed of eight parts. **Best viewed in color.**

2. ALGORITHM DESCRIPTION

This paper presents a novel approach to encode low-level descriptors into mid-level features using body structure information. As depicted in Fig.1, feature representation based on body-structure is performed after feature extraction, followed with KISSME-based person matching [11]. For body-structure based feature representation, human body is first split into eight parts according to body structure information [12] to construct the body-structure pyramid which is used as reference information to learn the body-structure codebook and pool the features encoded by LLC.

2.1. Body-structure pyramid

The substantial body structure information of pedestrian is very helpful for re-identification. Our body-structure pyramid is designed based on the following three observations: (1) Vertical space misalignments caused by pose and viewpoint variations appear much less than horizontal space misalignments; (2) Human body is not a rigid object for its complex kinematics, so it can be better described using a part-based model; (3) Spatial layout information is considerably critical information and can be used to describe body appearance.

Motivated by [5, 12], body structure information is utilized to construct the body-structure pyramid as shown in Fig.2. Adaptive part models based on background substraction techniques are used in some previous works [2, 13] and have gained some performance enhancements in certain situations. However, these adaptive part models require more computation and may generate incorrect segmentation in complex scenarios. In this paper, a simple but effective fixed part model is proposed to describe body appearance and solve the space misalignment by dividing the pedestrian image into increasingly fine vertical sub-regions with some prior knowledge of body structure. As depicted in Fig.2(b) three horizontal stripes of 16%, 29% and 55% of the total pedestrian height respectively locate head, torso and legs [12]. Further, torso part and leg part are both subdivided into two horizontal stripes with equal size as shown in Fig.2(c), so as to describe human in a finer level. The total eight parts in Fig.2(a)(b)(c) compose the body-structure pyramid.

2.2. Body-structure based feature representation

2.2.1. *Body-structure codebook learning*

In order to increase the discrimination of codebook, body-structure pyramid is used to construct the body-structure codebook. The patches, sampled from images, are divided into eight patch sets according to the body-structure pyramid:

$$P_a = \{p_{j,t} | p_{j,t} \in r_a, t = 1, ..., T\} \tag{1}$$

where P_a denotes the ath patch set and $p_{j,t}$ denotes the jth patch of the tth image, while r_a is the ath part of body-structure pyramid. K-Means is applied to learn the codebook using the descriptors extracted from patches randomly selected from the relevant patch set P_a. The final body-structure codebook consists of eight sub-codebooks as shown in Fig.1, and each sub-codebook has M entries with D dimensions:

$$\begin{aligned} B &= \{B_a | a = 1, ..., N\} \\ B_a &= [b_{a,1}, b_{a,2}, ..., b_{a,M}] \in \mathbb{R}^{D \times M} \end{aligned} \tag{2}$$

where B is the body-structure codebook and B_a is a sub-codebook. Parameter N denotes the number of sub-codebook and $N = 8$ in this paper.

2.2.2. *LLC coding*

In this paper, LLC [6] is adopted to encode mid-level features using body-structure codebook as shown in Fig.2. LLC gives an analytical solution for the following criteria:

$$\begin{gathered} \min_{C} \sum_{i=1}^{N} \|x_i - B_a c_i\|^2 + \lambda \|d_i \odot c_i\|^2 \\ s.t. 1^\top c_i = 1, \forall i \end{gathered} \tag{3}$$

where $\odot$ denotes the element-wise multiplication, and $d_i \in \mathbb{R}^M$ is a locality adaptor with different proportion for each basis. according to its similarity to the input descriptor x_i:

$$\begin{gathered} d_i = exp\left(\frac{dist(x_i, B_a)}{\sigma}\right) \\ dist(x_i, B_a) = [dist(x_i, b_{a,1}), ..., dist(x_i, b_{a,M})] \end{gathered} \tag{4}$$

where $dist(x_i, b_{a,j})$ is the Euclidean distance and σ is used to adjust the weight decay speed [6]. Regularization term in Eq.(3) leads to locality, which can generate similar codes for similar descriptors and make the feature more discriminative.

2.2.3. Body-structure pooling

Feature pooling is an effective way to select features and can achieve invariance of space misalignment. As shown in Fig.2, a novel feature pooling strategy using body-structure pyramid as reference information, is proposed in this paper. Body-structure pooling combines the codes of the same body part into a single feature vector and can make the feature vector invariant to space misalignment, especially the horizontal one caused by the varying poses and viewpoints. Since max pooling over sparse codes is robust to clutter [14] and can capture the salient properties of local regions [15], an operator is used here for feature pooling:

$$f_a = max(c_{a,1}, c_{a,2}, ..., c_{a,K}) \tag{5}$$

where "max" function runs in a row-wise manner, pooling codes $c_{a,i}$ in the ath part of body-structure pyramid into one feature vector f_a, and K denotes the number of descriptors in this part. Finally, feature representation is obtained by concatenating and normalizing the pooled features.

2.3. KISSME-based person matching

2.3.1. Distance calculation

Three low-level descriptors are used in our method, including (1) weighted HSV (wHSV) color histograms are extracted to capture color information as suggested in [2]; (2) SIFT descriptors are used to capture texture information and handle illumination variation; (3) LAB color histograms are extracted to enhance illumination invariance. The encoded wHSV, LAB, SIFT feature vectors are denoted as $wH(I)$, $LAB(I)$, $SIFT(I)$ respectively, and I is the pedestrian image.

Here, Mahalanobis distance is used to measure the distance between feature vector x_i and x_j:

$$d_M^2(x_i, x_j) = (x_i - x_j)^\top M (x_i - x_j) \tag{6}$$

where M denotes the Mahalanobis distance metric. In order to process large-scale person re-identification data, KISSME [11] is applied to learn the Mahalanobis distance metric. However, the feature vectors can't be processed by KISSME directly since their high-dimension may result in a singular matrix during matric learning. Generally, PCA is used to reduce feature dimension in most existing literatures [11, 16]. However, person re-identification data has large intra-class dissimilarity and inter-class similarity caused by variations of poses, illumination and viewpoints, resulted from which, LDA is more suitable than PCA in projecting re-identification data into the best identification space. Taking both effectiveness and speed into account, gaussian kernel PCA is first used to remove some noise and reduce dimension. Then LDA is used to further project features into the best identification low-dimension space.

2.3.2. Distance fusion

Since different features show different discrimination, a decision level fusion strategy is used to integrate the contributions of different features:

$$\begin{aligned} d_{BSFR}(I_A, I_B) = \beta_{wH} \cdot d_{wH}(wH(I_A), wH(I_B)) \\ + \beta_{LAB} \cdot d_{LAB}(LAB(I_A), LAB(I_B)) \\ + \beta_{SIFT} \cdot d_{SIFT}(SIFT(I_A), SIFT(I_B)) \end{aligned} \tag{7}$$

where d_{wH}, d_{LAB}, and d_{SIFT} are the normalized feature vector distances calculated by Eq.(6) while β_{wH}, β_{LAB}, β_{SIFT} denote the corresponding integrating weights.

3. EXPERIMENTS AND DISCUSSIONS

The detailed parameters are set as follows: images are divided into overlapping patches of size 8×8 with 4×4 stride. Each sub-codebook containing 1024 entities is constructed with 5000 patches randomly selected from the relevant patch set. The dimension of the feature vectors is reduced to 36.

3.1. Datasets and Evaluation Protocol

Our approach is evaluated on two publicly challenging datasets, VIPeR [17] and CUHK01 [18]. Experimental results are reported in the form of the average Cumulated Matching Characteristic (CMC) curve for 10 trials.

VIPeR dataset[1] contains 632 pedestrian image pairs which are taken from arbitrary viewpoints under varying illumination conditions and are normalized to 128×48 pixels. This dataset is randomly split into two parts, both consisting of 316 individuals, one for training and the other for testing. $\beta_{wH} = 2$, $\beta_{LAB} = 1$, $\beta_{SIFT} = 1$ is set for this dataset.

CUHK01 dataset[2] contains 971 individuals captured from two disjoint camera views. Each person has two images per camera view which are normalized to 160×60 pixels. This dataset is also split into two parts randomly. One contains 485 individuals for training, and the other contains 486 individuals for testing. As each person has two images in the probe and gallery respectively, the 4 distances between the image pairs are averaged to obtain the final distance. Here, $\beta_{wH} = 1$, $\beta_{LAB} = 1$, $\beta_{SIFT} = 1$, because higher image resolution may lead to more reliable SIFT descriptors.

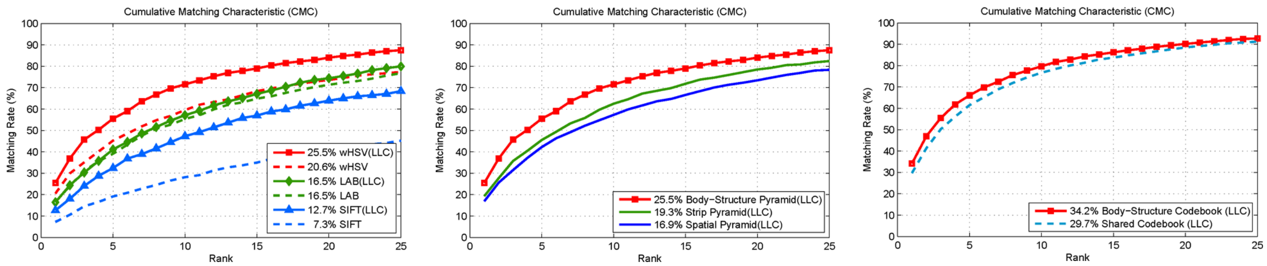
3.2. Evaluations and Analysis

The effectiveness of LLC coding strategy, body-structure pooling and body-structure codebook are all evaluated on VIPeR using KISSME to learn the mahalanobis matrix M.

Evaluation of LLC. LLC encodes low-level descriptors into mid-level features. Evaluation of LLC compares the performances using three low-level descriptors with and without LLC coding using shared codebook (*i.e.* the commonly used codebook in LLC [6]). All features are pooled via body-structure pooling. Fig.3(a) shows that for all the three low-level descriptors, performances using LLC are more competitive than that without using LLC. Take rank 10 for example, an improvement of 12.0% for wHSV and 19.2% for SIFT

[1] http://vision.soe.ucsc.edu/?q=node/178

[2] http://www.ee.cuhk.edu.hk/ xgwang/CUHK_identification.html

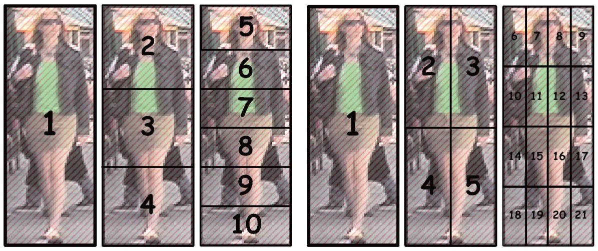


(a) Evaluation of LLC coding (b) Evaluation of body-structure pooling (c) Evaluation of body-structure codebook

Fig. 3. Evaluations on the VIPeR dataset. Rank-1 matching rate is marked before the name of each approach.

are respectively achieved over the original low-level descriptors. The main reason is that as an extension of BoF, LLC is good at handling space misalignment caused by different viewpoints and poses. Furthermore, the locality property of LLC can generate similar codes for similar descriptors, which may contribute to improving the discrimination of features.

Evaluation of body-structure pooling. To validate the effectiveness of body-structure pyramid on feature pooling, we compare the matching results of wHSV features pooled by three different spatial structures: body-structure pyramid, strip pyramid and spatial pyramid. Fig.4 depicts the detailed structures of strip pyramid and spatial pyramid. All features are encoded by LLC using shared codebook. Fig.3(b) shows that our proposed feature pooling by body-structure pyramid outperforms strip pyramid and spatial pyramid with an improvement of more than 5%. The reasonable explanation is that our proposed body-structure pyramid accords with human body structure better.



(a) strip pyramid (b) spatial pyramid

Fig. 4. Two compared pyramid structures.

Evaluation of body-structure codebook. To evaluate the effectiveness of body-structure codebook, wHSV, LAB and SIFT features encoded by LLC with body-structure pooling are employed. As depicted in Fig.3(c), body-structure codebook achieves better performance , since it can reflect characteristics of different body parts more accurately.

3.3. Comparison with state-of-the-arts

Comparing experiments of our BSFR and state-of-the-art approaches are conducted on VIPeR and CUHK01 datasets. Table 1 shows that BSFR clearly outperforms all the other state-of-the-art approaches on both datasets. To illustrate, BSFR has more competitive advantage over the latest approach SCNCD$_{all}$ [16] with best performances on VIPeR dataset and Mid-Filters [4] on CUHK01 dataset. The reasonable explanation is that BSFR makes best use of body structure information and uses mid-level features coded by LLC, which are insensitive to space misalignment and robust to variations of poses and viewpoints. In addition, our mid-level features are encoded using LLC with better discrimination and low computation complexity linear to the size of codebook and the number of the sampled patches.

Table 1. Comparisons with the state-of-the-arts on VIPeR

VIPeR	Rank 1	Rank 10	Rank 20	Rank 50
KISSME [11]	19.6	62.2	77	91.8
SDALF [2]	19.9	49.4	65.7	84.8
eBiCov [19]	20.7	56.2	68.0	-
Salience [4]	30.2	65.5	79.2	-
ARLTM [20]	21.2	38.7	52.9	67.5
Mid-Filters [9]	29.1	65.6	79.9	-
SCNCD$_{all}$ [16]	33.7	74.8	85.0	93.8
BSFR(Ours)	**34.2**	**79.7**	**90.2**	**98.1**

Table 2. Comparisons with the state-of-the-arts on CUHK01

CUHK01	Rank 1	Rank 10	Rank 20	Rank 50
SDALF [2]	9.9	30.3	41.0	-
ITML [9]	16.0	45.6	59.8	-
GenericMetric [18]	20.0	50.0	69.3	-
Salience [4]	28.5	55.7	68.0	-
Mid-Filters [9]	34.3	65.0	75.0	-
BSFR(Ours)	**36.2**	**72.2**	**83.4**	**93.5**

4. CONCLUSIONS AND FEATURE WORK

This paper introduces body-structure based feature representation (BSFR) for person re-identification. BSFR makes full use of body structure information by applying the novel body-structure pyramid in both codebook learning and feature pooling steps. Experimental results show that our approach can achieve better performance than state-of-the-art methods even in complex scenes with inter-class ambiguities. In addition, our method has low time consumption, which is applicable for the real-world surveillance systems. In future work, we plan to design a patch selection strategy in codebook learning step to enhance the discrimination, and use adaptive weights for feature fusion to adapt to different scenarios.

5. REFERENCES

[1] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, “The Re-Identification Challenge,” *in Person Re-Identification*, pp. 1–20, 2014.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, “Person re-identification by symmetry-driven accumulation of local features,” *in Proceedings of CVPR*, pp. 2360–2367, 2010.

[3] C. Liu, S. Gong, C. C. Loy, and X. Lin, “Person re-identification: what features are important?,” *in Proceedings of ECCV Workshops*, pp. 391–401, 2012.

[4] R. Zhao, W. Ouyang, and X. Wang, “Person re-identification by salience matching,” *in Proceedings of ICCV*, pp. 2528–2535, 2013.

[5] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” *in Proceedings of CVPR*, vol. 2, pp. 2169–2178, 2006.

[6] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” *in Proceedings of CVPR*, pp. 3360–3367, 2010.

[7] H. Liu, M. Liu, and Q. Sun, “Learning directional co-occurrence for human action classification,” *in Proceedings of ICASSP*, pp. 1235–1239, 2014.

[8] Q. Sun and H. Liu, “Action disambiguation analysis using normalized Google-like distance correlogram,” *in Proceedings of ACCV*, pp. 425–437, 2012.

[9] R. Zhao, W. Ouyang, and X. Wang, “Learning Mid-level Filters for Person Re-identifiation,” *in Proceedings of CVPR*, pp. 144–151, June 2014.

[10] Z. Yang, L. Jin, and D. Tao, “A comparative study of several feature extraction methods for person re-identification,” *in Proceedings of Biometric Recognition*, pp. 268–277, 2012.

[11] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, “Large scale metric learning from equivalence constraints,” *in Proceedings of CVPR*, pp. 2288–2295, 2012.

[12] S. Park and J. K. Aggarwal, “Simultaneous tracking of multiple body parts of interacting persons,” *in Computer Vision and Image Understanding*, vol. 102, no. 1, pp. 1–21, 2006.

[13] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, “Person Re-identification Using Haar-based and DCD-based Signature,” *in Proceedings of AVSS*, pp. 1–8, 2010.

[14] Y. L. Boureau, F. Bach, Y. LeCun, and J. Ponce, “Learning mid-level features for recognition,” *in Proceedings of CVPR*, pp. 2559–2566, 2010.

[15] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” *in Proceedings of CVPR*, pp. 1794–1801, 2009.

[16] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, “Salient Color Names for Person Re-identification,” *in Proceedings of ECCV*, pp. 536–551, 2014.

[17] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” *in Proceedings of PETS workshops*, 2007.

[18] W. Li, R. Zhao, and X. Wang, “Human Reidentification with Transferred Metric Learning,” *in Proceedings of ACCV*, pp. 31–44, 2012.

[19] B. Ma, Y. Su, and F. Jurie, “Bicov: a novel image representation for person re-identification and face verification,” *in Proceedings of BMVC*, pp. 57.1–57.11, 2012.

[20] X. Liu, M. Song, Q. Zhao, D. Tao, C. Chen, and J. Bu, “Attribute-restricted latent topic model for person re-identification,” *in Pattern Recognition*, vol. 45, no. 12, pp. 4204–4213, 2012.