Person Re-Identification using Pose-Driven Body Parts

Salwa Baabou[1,4], Behzad Mirmahboub[2], François Bremond[3], Mohamed Amine Farah[4] and Abdennaceur Kachouri[4]

[1] University of Gabes, National Engineering School of Gabes, Tunisia
[2] Pattern Analysis and Computer Vision (PAVIS), Italian Institute of Technology, Genova, Italy
[3] INRIA Sophia Antipolis Mediterranee, France
[4] University of Sfax , National Engineering school of Sfax Laboratory of Electronics and Information Technology (LETI), Sfax, Tunisia
baabousalwa@gmail.com

Abstract. The topic of Person Re-Identification (Re-ID) is currently attracting much interest from researchers due to the various possible applications such as behavior recognition, person tracking and safety purposes at public places. General approach is to extract discriminative color and texture features from images and calculate their distances as a measure of similarity. Most of the work consider whole body to extract descriptors. However, human body maybe occluded or seen from different views that prevent correct matching between persons. To this end, we propose in this paper to use a reliable pose estimation algorithm to extract meaningful body parts. Then, we extract descriptors from each part separately using LOcal Maximal Occurrence (LOMO) algorithm and Cross-view Quadratic Discriminant Analysis (XQDA) metric learning algorithm to compute the similarity. A comparison to some recent state-of-the-art Re-ID methods in most commonly used benchmark Re-ID datasets will be also presented in this work.

Keywords: Person Re-Identification (Re-ID), Pose-driven body parts, LOMO features, XQDA algorithm

1 Introduction

In the modern computer vision community, the emergence of person Re-Identification (Re-ID) is related to the increasing demand of public safety and the widespread of large camera networks. From this perspective, the task of person Re-ID consists in recognizing and identifying a person between several non overlapped camera views. The images in two cameras are called “probe” and “gallery” sets in which we are looking for probe images between gallery images. It has important applications in surveillance systems and can reduce human labor and errors of human matching.

However, the most challenging problem of Re-ID is how to correctly match two images of the same person under intensive appearance changes, such as lighting, pose and viewpoint changes.The problem of localizing keypoints or parts of human body is known as human pose estimation which consists in finding or extracting body parts of individuals. However, this task presents a set of challenges by its own: *i)* an image may contain one or more persons that can occur at any time or position. *ii)* this interaction between persons may lead to complex interference which make association of parts difficult.*iii)* making realtime performance is a challenge due to

runtime complexity which increases with the number of persons present in the image or scene; *i.e* the more people there are, the greater the computational cost is.
In the literature, there are many approaches that focus to extract body parts of individuals [1, 2, 3].

To sum up, we outline the main contribution of this paper as follows: Extracting body parts from iLIDS-VID [10], PRID-2011 [11], MARS [12] and our own dataset called CHU-Nice using OpenPose [6] which is a state-of-the-art pose estimation algorithm that detects 15 body joints. From these body parts, we extract LOcal Maximal Occurrence (LOMO) features and then compute the similarity using Cross-view Quadratic Discriminant (XQDA) algorithm [4]. Then, we compare our work to some recent state-of-the-art Re-ID approaches.

This paper is organized as follows: Section II is the core of the paper: it presents body parts extraction: We extract the features from those body parts using LOMO method and then we compute the distance using XQDA algorithm between those descriptors. In section III, we present some commonly used Re-ID benchmark datasets that we will use to evaluate our approach and we compare it to some recent state-of-the-art Re-ID approaches. Finally, we finish by drawing the conclusion.

2 State-of-the-art Re-ID methods

Person Re-Identification (Re-ID) is the problem of identifying and recognizing a person between several non-overlapped camera views at different times and locations.

Person Re-ID methods are categorized into either: *i)* signature modeling-based methods, by extracting discriminative characteristics from the different body parts, or *ii)* matching function learning-based approaches, by optimizing the parameters of the similarity function in order to minimize intra-class variance and maximize inter-class variance of signatures.

Moreover, authors in [16] proposed a division that exploits the anti-symmetry property of clothes structure that divides the body into three regions (head, torso and legs). Besides, in [15], authors presented a pre-learned model to detect the different body parts disposition. However, these models suffer from the large viewpoint variation and partially occluded persons. Khan et al. [14], divide the body into three stripes corresponding to the head, torso and legs with size of 16, 29 and 55 respectively. The most widely used division method [7] presents six stripes of the human body which are the head, upper and lower torso, upper and lower legs, and feet.

Most of these methods consider that all the different body parts are uniformly informative for appearance modeling.

3 Proposed Method

Fig.1 illustrates our proposed framework which consists of five steps. We begin by detecting the body joints of persons using pose estimation algorithm. Then, we extract the body parts. From these latter, we extract LOcal Maximal Occurrence (LOMO) features and we evaluate the similarity by computing the distances between

the descriptors extracted by using the Cross-view Quadratic Discriminant (XQDA) metric learning algorithm.

Fig.1. Overview of our proposed framework

3.1 Body Parts extraction

OpenPose [5] is a state-of-the-art pose estimation algorithm that detects 15 body joints as shown in Fig.1 (a). An example of pose estimation result on MARS dataset [12] is shown in Fig.1(b). We used joint positions to define 12 body parts as shown in Fig.1(c). Our idea is to extract image descriptor from each part and calculate their distances separately. Distance between two images can be computed by weighted average of all distances between body parts.

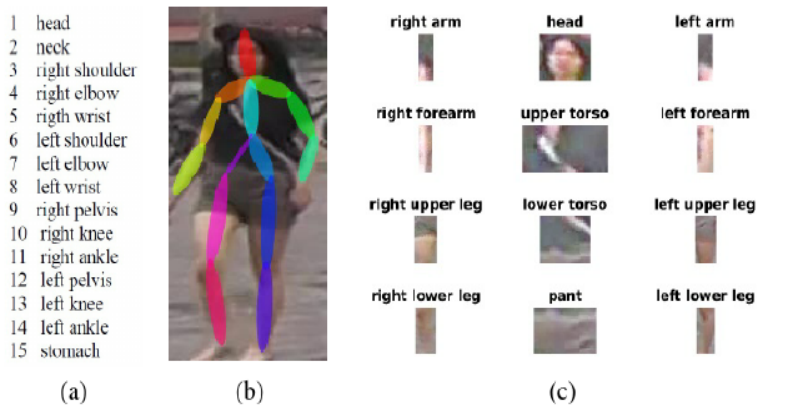

Fig. 2. Human body joints and parts (a) Body joints that are detected withpose estimation algorithm (b) An example image from MARS dataset withestimated pose (c) Different body parts that we defined for feature extraction.

LOMO [4] is a famous descriptor for person Re-Identification that divides each image into horizontal stripes and finds the maximum bins of color and texture histograms in each stripe. We modified this code to use it on body parts.

After extracting LOMO features, next step is to compare probe feature vector x_i with gallery feature vector x_j, find their similarity and calculate their distances in order to find a correct match between gallery and probe images. Different metric learning methods are proposed in literature but in our work we will use the XQDA metric learning algorithm.

3.2 LOMO feature extraction

LOcal Maximal Occurrence (LOMO) [4] is a state-of-the-art feature extraction method that aims to compensate illumination variations and viewpoint changes between two cameras. In fact, by applying Retinex algorithm [18], it pre-processes person images in order to produce a color image that is consistent to human observation of the scene especially in shadowed regions.

Fig.3(a) shows an example of original and processed images of the same person across two cameras. Using Retinex images makes person Re-ID easier than using the original images. After adjusting the illumination, HSV color histogram and Scale Invariant Local Ternary Pattern (SILTP) texture descriptor are extracted from images. SILTP is an improved operator over Local Binary Pattern (LBP) that aims to achieve invariance to intensity scale changes and robustness to image noises.

Fig.3(b) shows the LOMO scheme to address view point changes between two cameras. A sliding window with size of 10x10, with an overlapping step of 5 pixels, locates local patches in 128x48 images. All sub-windows at the same horizontal location are checked and maximum values between all corresponding bins are selected to produce only one histogram for each row. The above feature extraction procedure is repeated for two additional scales by down sampling the original image using 2x2 local average pooling operations (that produces 64x24 and 32x12 image size) to consider the multi-scale information. All the computed local maximal occurrences are concatenated to get the final descriptor which has (8x8x8 color bins + 3^4 x2 SILTP bins)x(24+11+5 horizontal groups) = 26,960 dimensions.

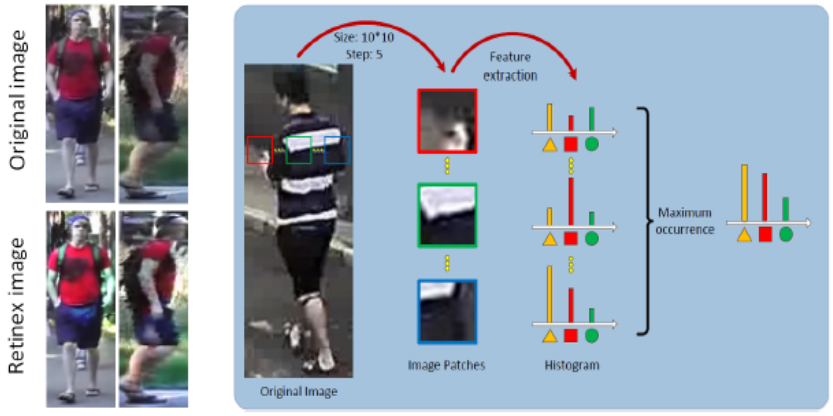

Fig. 3. Illustration of the LOMO feature extraction method [14]

3.3 XQDA algorithm

Usually the original feature dimension is large and a low dimensional space is preferred for classification.

Authors in [4] proposed Cross-view Quadratic Discriminant Analysis (XQDA) algorithm that is an extension of KISSME method [17]. The idea is to learn a lower dimensional subspace W that original features are mapped to it and at the same time learn a distance function in that subspace for the cross-view (different camera) similarity measure. For this purpose, the distance function is modified as:

$$d_w(x_i, x_j) = (x_i - x_j)^T W(\sum\nolimits_I^{'-1} - \sum\nolimits_E^{'-1})W^T(x_i - x_j) \qquad (1)$$

Where $\sum_I^{'} = W^T \sum_I W$ and $\sum_E^{'} = W^T \sum_E W$.

Therefore, we need to learn a kernel matrix $M_w = W(\sum_I^{'-1} - \sum_E^{'-1})W^T$. On the other hand, the goal is to find a projection direction w that increase extra-personal variance $\sigma_E(w)$ and decrease intra-personal variance $\sigma_I(w)$. Since $\sigma_E(w) = w^T \sum_E w$ and $\sigma_I(w) = w^T \sum_I w$, then maximizing the objective $\frac{\sigma_E(w)}{\sigma_I(w)}$ corresponds to the Generalized Rayleigh Quotient:

$$J(w) = \frac{w^T \sum_E w}{w^T \sum_I w} \qquad (2)$$

That is, the largest eigenvalue of $\sum_I^{-1} \sum_E$ is the maximum value of $J(w)$, and the corresponding eigenvector w_1 is the solution.

4 Datasets and Performance Evaluation

4.1 Datasets

The commonly used datasets for image- and video-based person Re-ID are summarized in Table1 [6]. We have evaluated our work on four challenging benchmark datasets: *PRID-2011, iLIDS-VID*, *MARS* and on our own dataset *CHU-Nice dataset*.

- **PRID-2011**[11] is a multi-shot dataset captured by two cameras in outdoor environment. Camera A captures 385 persons and camera B captures 749 persons. Only the first 200 persons are common between two cameras. Each person in each camera has 5-675 consecutive frames.
- **iLIDS-VID** [10**]** consists of 300 identities and each identity has 2 image sequences, totaling 600 sequences. The length of image sequences varies from 23 to 192, with an average number of 73. This dataset is more challenging due to environment variations. The test and training set both have 150 identities.

- **MARS** [12] (Motion Analysis and Re-identification Set) is an extension of Market1501 dataset. It contains 1261 persons with about 1.19 million images and 3248 distractors.
- **CHU Nice** dataset is collected in the hospital of Nice (CHU) in Nice, France. It is related to INRIA Sophia Antipolis. Mostof the people recruited for this dataset were elderly people, aged 65 and above, of both genders. It contains 615 videos with 149365 frames. It's also an RGB-D dataset, *i.e* it provides RGB+Depth images.

Table 1. Summary of some widely used datasets from image- and video-based Person Re-ID [6]

Datasets	#ID	#Image	#Distractors	#Camera
ViPER	632	1,264	0	2
iLIDS	119	476	0	2
GRID	1025	1,275	775	8
CAVIAR	72	610	22	2
PRID2011	934	1,134	732	2
CUHK01	971	3,884	0	2
CUHK02	1,816	7,264	0	10 (5pairs)
CUHK03	1,467	13,164	0	10 (5pairs)
RAiD	43	1,264	0	4
PRID450S	450	900	0	2
Market-1501	1,501	32,668	0	6
ETHZ	148	148	0	1
3DPES	192	1,000	0	8
iLIDS-VID	300	600	0	2
MARS	1261	20,715	0	6
DukeMTMC-reID	1,812	36,441	408	8
DukeMTMC4ReID	1,852	46,261	439	8

4.2 Performance Evaluation

The Cumulative Matching Characteristic (CMC) curve is the metric adopted where each element in the gallery is ranked based on its comparison to the probe. The probability that the correct match is ranked equal to or less than a particular value is plotted against the size of the gallery set. However, when multiple ground truth exist in the gallery and inspired from the assumption that a perfect Re-ID system should be able to return all true matches to the user, the mean Average Precision (mAP) is proposed for evaluation. This latter allows to know whether most of the matched gallery images have been ranked high in the output of Re-Identification ranking or not. In the case of the Market-1501 dataset, mAP and CMC are used together for evaluation where multiple ground truths exist for each query from multiple cameras.

In our case, we used the CMC and mAP as evaluation metrics for our experiments.

In Table 2, we compare our results ([5]+LOMO+XQDA [4]) with some state-of-the-art methods in the context of video-based datasets (iLIDS-VID, PRID-2011,MARS and CHU-Nice) as we are trying to propose our new dataset CHU-Nice which is a multi-shot Re-ID dataset. Three descriptors are compared, *i.e*BoW [7], HOG3D [8], LOMO with the metric learning algorithm XQDA [4] which is evaluated.
From the above results, we notice that our proposed method achieves the best Rank-1 accuracy 54.8% in the four datasets (for example: Rank-1 accuracy =54.8% on iLIDS-VID dataset). However, we believe that the research on both image- and video-based Re-ID still has good potential for improvement in the future especially with the emergenceof large-scale datasets and the great success of deep Convolutional Neural Network CNN system in computer vision which have made great influence in person Re-ID.

Table 2. Rank-1 accuracy (%) with comparison to some state-of-the-art methods on the four datasets is presented . We use average pooling for iLIDS-VID, and max-pooling for PRID-2011, MARS and CHU Nice datasets. Best results are highlighted in bold.

Methods	iLIDS-VID	PRID-2011	MARS	CHU-Nice
BoW+XQDA[5,8]	14.0	31.8	30.6	-
HOG3D+XQDA[5,9]	16.1	21.7	2.6	-
LOMO+XQDA[14]	53.0	-	30.7	36.2
Ours ([6]+LOMO+XQDA[5])	**54.8**	**-**	**32.7**	**40.6**

5 Conclusion

In this paper, we proposed to use a reliable pose estimation algorithm to extract meaningful body parts and then extract LOMO descriptors from each part separately and then compute the distances between those descriptors using XQDA metric learning algorithm as a measure of similarity.
Preliminary experiments show some potentials of using pose estimation for Re-ID, but not as accurate as global signature.

One shortcoming of our work may be that we relied on LOMO descriptor that is essentially designed for the whole image. Suitable descriptor such as deep features should be designed for body parts. In case of proper descriptor, part-based Re-Identification is promising to cope with the problem of pose and viewpoint variations. This work can also be extended to detect mid-level features or attributes (such as gender, long hair, jeans, t-shirt etc.) that are more reliable than low-level descriptors (such as gradients and histogram).

References

1. A. Bulat and G. Tzimiropoulos.: Human pose estimation via convolutional part heatmap regression. In ECCV, (2016).
2. V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh.: Pose machines: Articulated pose estimation via inference machines. InECCV, (2014).
3. S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. : Convolutional posemachines. In CVPR, (2016).
4. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR, (2015).
5. Z. CAO, T. SIMON, S.-E. WEI, Y. SHEIKH. Realtime Multi-Person 2DPose Estimation using Part Affinity Fields, in ”CVPR”, (2017).
6. M. Gou. Person re-identification datasets. http://robustsystems.coe.neu.edu/sites/ robustsystems.coe.neu.edu/files/systems/projectpages/reiddataset.html, (2017).
7. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person reidentification: A benchmark. In: CVPR,(2015).
8. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3dgradients. In: BMVC, (2008).
9. Fendri, E., Frikha, M., &Hammami, M.: Adaptive Person Re-identification Based on Visible Salient Body Parts in Large Camera Network. The Computer Journal, 60(11), 1590-1608, (2017).
10. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by videoranking. In: Computer VisionECCV 2014, pp. 688703. Springer, (2014).
11. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identificationby descriptive and discriminative classification. In: Image Analysis, pp.91102, (2011).
12. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q. Mars: A video benchmark for large-scale person re-identification. In EuropeanConference on Computer Vision, ECCV , pp. 868-884, Springer, (2016).
13. Cheng, D.S. and Cristani, M. : Person re-identification by articulated appearance matching. In Person Re-Identification, 139–160. Springer, (2014).
14. Khan, A., Zhang, J. and Wang, Y. : Appearance-based Re-identification of People in Video. 2010 Int. Conf. Digital Image Computing: Techniques and Applications (DICTA), pp. 357–362. IEEE, (2010).
15. Jaouedi, N., Boujnah, N., Htiwich, O., &Bouhlel, M. S.: Human action recognition to human behavior analysis. In 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT) , pp. 263-266. IEEE, (2016).
16. Das, A., Chakraborty, A. and Roy-Chowdhury, A.K.: Consistent Re-identification in a Camera Network. European Conf. Comput. Vision, pp. 330–345. Springer, (2014).
17. M. Koestinger, M. Hirzer, P.Wohlhart, P. M. Roth, and H. Bischof. Largescale metric learning from equivalence constraints. In IEEE Conferenceon Computer Vision and Pattern Recognition (CVPR), pages 22882295.IEEE, (2012).
18. D. J. Jobson, Z.-u. Rahman, and G. A. Woodell. A multiscale retinex for bridging the gap between color images and the human observation ofscenes. IEEE Transactions on Image processing, 6(7):965976, (1997).