

Images Selection and Best Descriptor Combination for Multi-shot Person Re-identification

Yousra Hadj Hassen^(✉), Kais Loukil, Tarek Ouni, and Mohamed Jallouli

National School of Engineers of Sfax, Computer and Embedded Systems Laboratory,
University of Sfax, 4.7 km Street of Soukra, 3038 Sfax, Tunisia

hadjhassen.yousra@gmail.com

<http://www.ceslab.org>

<http://www.enis.rnu.tn>

Abstract. To re-identify a person is to check if he/she has been already seen over a cameras network. Recently, re-identifying people over large public cameras networks has become a crucial task of great importance to ensure public security. The vision community has deeply studied this area of research. Most existing researches rely only on the spatial appearance information extracted from either one (single-shot) or multiple images (multi-shot) for each person. Actually, the real person re-identification framework is a multi-shot scenario. However, to efficiently model a person's appearance and to select the most informative samples remain a challenging problem. In this work, an extensive comparison of descriptors of state of art associated to the proposed frame selection method is considered. Specifically, we evaluate the samples selection approach using different known descriptors. For fair comparisons, two standard datasets PRID 2011 and iLIDS-VID are used showing the effectiveness and advantages of the proposed method.

Keywords: Camera network · Descriptor · Model · Multi-shot · Person re-identification · Selection

1 Introduction

Recently, person re-identification (re-id) in non-overlapped cameras network presents a crucial task for many real applications like video surveillance, multimedia applications, behavior recognition,... [1]. To re-identify a person is to match his identity across camera views despite the changes that may occur. Actually, many environmental constraints can alter a persons appearance over different cameras views such as luminance variations, different point of view, scale zooming as shown in Fig. 1.

The proposed approaches can be classified either as single-shot or multiple-shot methods depending on the number of images used to construct the person identity. Contrary to the use of a single image to re-identify a person, the multi-shot based methods have been largely studied and significant results are achieved

[2]. They mainly focus on designing discriminative feature descriptors collected over many images of the same person. Whereas, real objectives of person re-id steel far from being reached because both relatively reduced execution time and robust feature descriptor are required.

In this paper, the trade-off between the use of robust descriptor and the reduction of execution time is considered. Multiple-shot proposed methods give potential results for person re-id while ignoring the selection of shots used for the re-id. Actually, most of multiple-shot methods select randomly the images forming the identity of the person but try to generate sufficiently robust descriptors that handle appearance changes caused by scale, lighting variations, view angles conditions and occlusions (Fig. 1). Since, for the multi-shot case, the results of images selection methods have a strong impact not only on the descriptor used for re-id, but also on the overall processing time of the system, the selection of both robust descriptor and discriminative frames to construct representative identity for each person is studied.

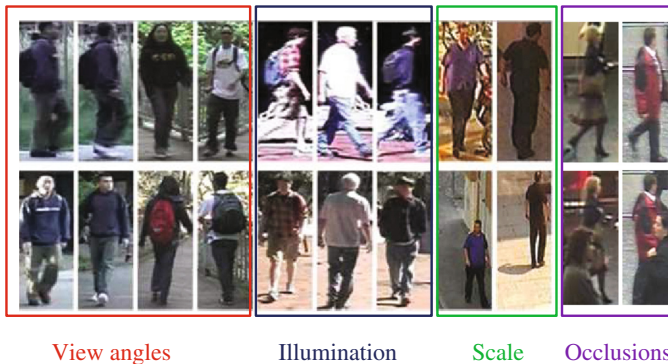


Fig. 1. Re-identification system constraints.

A key frame selection method is proposed and a rich comparison of recent robust proposed descriptors is associated to find the most performing combination for a real person re-id system.

The paper is organized in 5 sections. The first section is the introduction, followed by Sect. 2 of the related work. Section 3 describes the overall framework. The evaluation and the results comparison are detailed in Sect. 4. The final section is the conclusion and the future works.

2 Related Work

The contribution of this work may be considered on two fields; person re-identification descriptors as well as samples selection methods.

2.1 Descriptors

In person description, the most commonly used features are color and texture. Symmetry Driven Accumulation of Local Features (SDALF) [3] exploits the symmetric property of a person through obtaining head, torso, and leg positions to handle view variations. Gheissari et al. [4] propose a spatial-temporal segmentation method to detect stable foreground regions. For a local region, an HS histogram and an edgel histogram are computed. The latter encodes the dominant local boundary orientation and the RGB ratios on either sides of the edgel.

Hirzer et al. propose a generic descriptive statistical model in [5] and the appearance is modeled by a set of region covariance descriptors. Gray and Tao [6] use 8 color channels (RGB, HS, and YCbCr) and 21 texture filters on the luminance channel, and the pedestrian is partitioned into horizontal stripes. Similarly, Mignon et al. [7] build the feature vector from RGB, YUV and HSV channels and the LBP texture histograms in horizontal stripes. In [8–12], the 32-dim LAB color histogram and the 128-dim SIFT descriptor are extracted from each 10×10 patch densely sampled with a step size of 5 pixels. Das et al. [13] apply HSV histograms on the head, torso and legs from the silhouette. Pedagadi et al. [14] extract color histograms and moments from HSV and YUV spaces before dimension reduction using PCA. Liu et al. [15] extract the HSV histogram, gradient histogram and the LBP histogram for each local patch. In [16], Liao et al. propose the local maximal occurrence (LOMO) descriptor, which includes the color and SILTP histograms. Bins in the same horizontal stripe undergo max pooling and a three-scale pyramid model is built before a log transformation.

Ayedi et al. propose a multi-scale covariance descriptor in [17] using a quad-tree feature to tackle scale zooming appearance and occlusions in person re-id. In [18], Zheng et al. propose extracting the 11-dim color names descriptor for each local patch, and aggregating them into a global vector through a Bag-of-Words (BoW) model. In [19], a hierarchical Gaussian feature is proposed to describe color and texture cues, which models each region by multiple Gaussian distributions. Each distribution represents a patch inside the region. Liu et al. [20] improve the latent Dirichlet allocation (LDA) model using annotated attributes to filter out noisy LDA topics. In [21], Su et al. embed the binary semantic attributes of the same person but different cameras into a continuous low-rank attribute space, so that the attribute vector is more discriminative for matching. Shi et al. [22] propose learning a number of attributes including color, texture, and category labels from existing fashion photography. In [23], where the image is divided into a number of subblocks, each with its associated color histogram, multi-precision similarity matching is granted despite scale and lighting variations.

2.2 Sample Selection

Video summarization is the most known field in representative selection applications. The problem has been treated using clustering, vector quantization [24–26]

or sparse selection [27–29]. In [29], the Sparse Modeling Representative Selection (SMRS), based on the summary of videos by considering the proximity of the selected frames in the timeline, removes redundant frames. Intra and inter iteration redundancy splitting is proposed in [30] and significant re-id results are achieved.

Most of proposed representative samples selection methods rely on appearance variation in time, however, time information is almost unavailable in person re-id datasets. The proposed framework, based on key clusters and key frames selection, takes care of this issue. This enables the selection of as many informative samples as possible to improve the identification performance but at the same time avoids tedious training and useless gallery images. Multi-shot are outperforming single-shot re-id approaches. Incorporating supervision using training data leads to superior performance, which is the goal of metric learning. However, for a multi-shot metric learning person re-id approach with superior real scenario performance, the amount of data presents a hard issue for training time. That's why, we propose a multi-shot metric learning person re-id framework based on continuous representative samples selection.

3 Proposed Approach

In Fig. 2, the input of the proposed framework is a large gallery of images mostly formed by large sequence of frames of different persons captured in one camera view.

The proposed overall scheme of person re-id is composed of three main parts; first, features are extracted from the images in order to project the visual appearance into concrete parameters (color, texture, position, pose view). After that, a key frame selection algorithm is introduced [31]. The goal is to keep only informative images for each person so that all the informative appearance variations over time and space are summarized and useless noisy and redundant frames are removed. The major contribution of this work is to get the best combination of the descriptor (set of features) and the key frame selection algorithm so that a relatively speed and robust (accurate) re-id is achieved. To that end, the inter and intra descriptors distances are computed. The feed-back enables the evaluation of the used features and the algorithm converges to the best (descriptor, key frame selection) combination. Thus, a new gallery is formed containing key frames descriptors. Finally, the matching block allows the mapping of a newly unknown observed individual in another camera view (a probe) to one of the identity stored in the gallery yet constructed.

3.1 Features Extraction

As detailed in section two, different descriptors are proposed and excellent performances are reached. It is pretty promising to construct robust identities for each person. However, execution time and memory consumption present crucial constraints that must be treated for real video surveillance applications. Therefore,

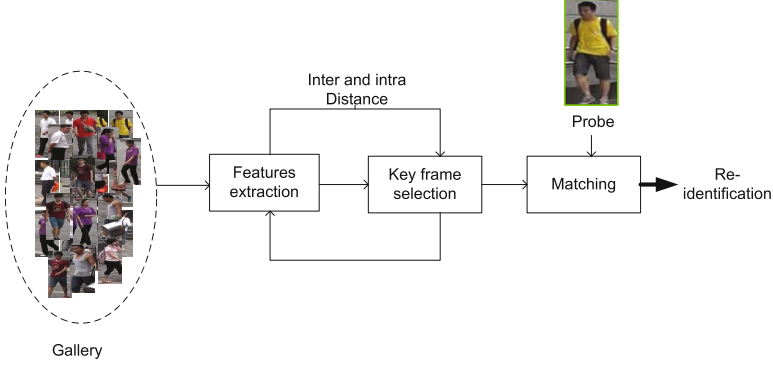


Fig. 2. Overall proposed approach.

person re-id systems have to take into account the trade-off between memory consumption (and so time consumption) and accuracy. Generally, robust descriptors can not be applied for a practical scenario due to their complex mathematical principals and large memory enquiries. So, basic descriptors such as covariance [5] treating lighting variations constraint, HOG [23] dealing with color and pose variations and multi-scale covariance [17] studying the scale zooming and occlusions, are tested. These descriptors are complementary in treating the different re-id system challenging issues. The impact of the key frame selection algorithm is evaluated on both re-id rates and memory gain.

3.2 Key Frame Selection

To concisely summarize long sequences of moving persons in one camera view (tracks), a representative frames selection method is proposed in Fig. 3.

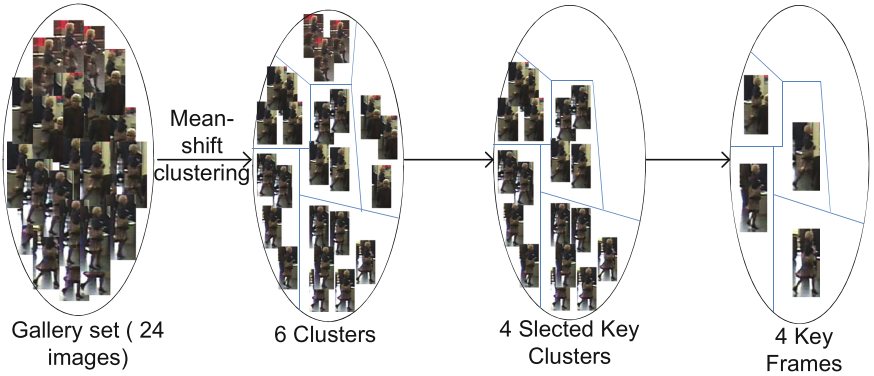


Fig. 3. Detailed example of key frame selection for person 1 camera 1 of PRID_2011.

Since redundancy presents a hard issue for most of person re-id systems, the use of a clustering algorithm may be a performing solution to surpass such problem. To that end, the mean shift clustering algorithm is used to form groups of similar images; as shown in Fig. 3, 24 different images describe 6 appropriate visual appearances. In Fig. 3, 2 useless clusters are removed. Then, non-frequent frames forming small clusters, in term of size or frames number, are removed and only informative clusters are kept. A representative frame is selected as a head, for each yet selected key cluster, so that it replaces the whole cluster in the person identity. As in the example presented in Fig. 3, instead of extract features of 24 images, only 4 images will form the person identity. Algorithm 1 details the key frame selection algorithm and conditions of the choice of key clusters or key frames are defined.

Algorithm1: Key frame selection based clustering

Input: set of N target captures

Output: set of discriminative target captures

Initialization: set radius for mean shift clustering

```

1: Obtain a number of clusters by performing
   the mean-shift clustering process in the
   feature space among the N samples.
2: Find the key cluster such;
   Size (key cluster) > Mean (Size (clusters))/2
3: Find the key frame such;
   Distance (key frame, center cluster) is minimal.

```

3.3 Matching

To re-identify an unknown newly captured person is to match his image to an identity stored in the gallery. To that end, a multi-class SVM classifier, yet trained by the gallery set, is tested for the query image. So, the SVM map the unknown person to the most similar trained identity. As detailed in Fig. 4, the tracked persons in one camera view are modeled and their identity based on key frames are stored and fed, in a first step, into the SVM through a file ‘train’ representing the gallery. In a second step, the descriptor of a query image, saved in a ‘test’ file, is mapped to one of the classes of the gallery by the SVM. The output of the matching phase is a label that re-identify the tested image.

4 Evaluation

4.1 Datasets

For experiments, we use multi-shot standard datasets PRID_2011 [32] and iLIDS-VID [33]. These two datasets are very challenging due to clothing similarities among people, lighting and view point variations across camera views, cluttered background and occlusions.

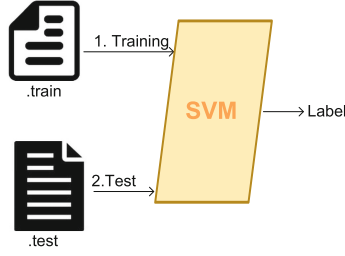


Fig. 4. Matching based SVM classifier.

PRID.2011 dataset: PRID.2011 dataset is formed of images of 200 and 749 people captured by two cameras A and B respectively. Each person has 5 to 675 images available. This dataset is hard because it presents real images with noisy background and illumination variations.

iLIDS-VID dataset: it presents different frames of 300 people captured by two non-overlapped cameras in an airport arrival hall. It is a challenging dataset due to the huge amount of images per person with clothing similarities and both partial and total occlusions.

4.2 Test Design

The results of re-id rates and memory consumption are computed in Tables 1 and 2 for respectively PRID.2011 and iLIDS-VID datasets for the three basic descriptors covariance [5], histogram of gradient (HOG) [23] and multi-scale covariance [17]. Of course, the person re-id rates are not very promising using the proposed discriminative selection but they still similar to the competing rates given by the robust descriptors in the full training case i.e. using the whole dataset for training. Moreover, the memory gain is significant thanks to the notable reduction of the trained data for the SVM classifier. Thus, real world re-id scenarios could be efficiently treated.

Table 1. Memory gain and re-id rate for 3 descriptors for PRID.2011

Descriptors	Re-id rates		Memory gain
	Full training	Key frame training	
Covariance	84.2%	74.3%	80%
HOG	56.7%	43.1%	95%
Multi-scale covariance	97%	72.8%	96%

Table 2. Memory gain and re-id rate for 3 descriptors for iLIDS-VID

Descriptors	Re-id rates		Memory gain
	Full training	Key frame training	
Covariance	84.9%	78.1%	88%
HOG	50.2%	48.5%	97%
Multi-scale covariance	81.9%	72.9%	94%

4.3 Results Discussion

The three basic descriptors compared are frequently used in person re-id. The results shown in Tables 1 and 2 demonstrate that the covariance descriptor outperforms HOG and Multi-scale covariance descriptors for both tested datasets. Thanks to the robust extracted features for the covariance descriptor, mainly colors and texture, the reached re-id rate is about 84%. Actually, it is an efficient result suitable for a practical real video surveillance application. The evaluation proves that to model some representative frames by a selected robust descriptor seems to be a promising solution to guaranty both re-id efficiency and memory consumption gain. However, a real world person re-id scenario steels far from being solved. In fact, the use of the robust classifier SVM shows that training key frames significantly outperforms full training (i.e. train all the images available for each person) in terms of execution time and memory consumption. However, the re-id rate steel greater for the latter case. Thats why, an extensive comparison of different classifier (as Adaboost) and the impact in re-id results may be proposed in future works.

5 Conclusion

Person re-id, has become an inherently task, of extensive interest, for real video surveillance applications. Having proposed a novel algorithm of representative images selection, the goal, in this paper, is to associate the most robust descriptor giving best re-id results. A comparative analysis of frequently used descriptors for re-id frameworks is conducted by computing the re-id rate. The impact of the key frames selection algorithm is highlighted thanks to the significant reduction of memory consumption and of course execution time. Finally, the combination of a selected descriptor with selected frames leads to an efficient multi-shot person re-id system. The proposed approach will be deeply studied to reach more performing real-world results and further evaluations will be delivered in future works.

References

1. Karanam, S., Gou, M., Wu, Z., Rates-Borras, A., Camps, O., Radke, R.J.: A comprehensive evaluation and benchmark for person re-identification: features, metrics, and datasets. arXiv preprint [arXiv:1605.09653](https://arxiv.org/abs/1605.09653) (2016)

2. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future. arXiv preprint [arXiv:1610.02984](https://arxiv.org/abs/1610.02984) (2016)
3. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. *Comput. Vis. Image Underst.* **117**, 130–144 (2013)
4. Gheissari, N., Sebastian, T.B., Hartley, R.: Person re-identification using spatiotemporal appearance. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1528–1535. IEEE Press, New York (2006)
5. Hirzer, M., Belezni, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian Conference on Image Analysis, pp. 91–102. Springer, Heidelberg (2011)
6. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision, pp. 262–275. Springer, Marseille (2008)
7. Mignon, A., Jurie, F.: PCCA: a new approach for distance learning from sparse pairwise constraints. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2666–2672. IEEE Press, Providence (2012)
8. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3586–3593. IEEE Press, Portland (2013)
9. Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L., Smith, J.R.: Learning locally-adaptive decision functions for person verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3610–3617. IEEE Press, Portland (2013)
10. Chen, D., Yuan, Z., Chen, B., Zheng, N.: Similarity learning with spatial constraints for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1268–1277. IEEE Press, Las Vegas (2016)
11. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2528–2535. IEEE Press, Sydney (2013)
12. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 144–151. IEEE Press, Columbus (2014)
13. Das, A., Chakraborty, A., Roy-Chowdhury, A.K.: Consistent re-identification in a camera network. In: European Conference on Computer Vision, pp. 330–345. Springer, Zurich (2014)
14. Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B.: Local fisher discriminant analysis for pedestrian re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3318–3325. IEEE Press, Portland (2013)
15. Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., Bu, J.: Semi-supervised coupled dictionary learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3550–3557. IEEE Press, Columbus (2014)
16. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206. IEEE Press, Boston (2015)
17. Ayedi, W., Snoussi, H., Abid, M.: A fast multi-scale covariance descriptor for object re-identification. *Pattern Recogn. Lett.* **33**, 1902–1907 (2012)

18. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1116–1124. IEEE Press, Chile (2015)
19. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical gaussian descriptor for person re-identification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1363–1372. IEEE Press, Las Vegas (2016)
20. Liu, X., Song, M., Zhao, Q., Tao, D., Chen, C., Bu, J.: Attribute restricted latent topic model for person re-identification. *Pattern Recogn.* **45**, 4204–4213 (2012)
21. Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L.S., Gao, W.: Multi-task learning with low rank attribute embedding for person re-identification. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3739–3747. IEEE Press, Chile (2015)
22. Shi, Z., Hospedales, T.M., Xiang, T.: Transferring a semantic representation for person re-identification and search. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4184–4193. IEEE Press, Boston (2015)
23. Lin, S., Ozsus, M.T., Oria, V., Ng, R.: An extensible hash for multi-precision similarity querying of image databases. In: *Proceedings of the 27th International Conference on Very Large Databases, Italy*, pp. 221–230 (2001)
24. De Avila, S.E.F., Lopes, A.P.B., da Luz, A., de Albuquerque Arajo, A.: VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recogn. Lett.* **32**, 56–68 (2011)
25. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. *Science* **315**, 972–976 (2007)
26. Garcia, S., Derrac, J., Cano, J., Herrera, F.: Prototype selection for nearest neighbor classification: taxonomy and empirical study. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 417–435 (2012)
27. Cong, Y., Yuan, J., Luo, J.: Towards scalable summarization of consumer videos via sparse dictionary selection. *IEEE Trans. Multimedia* **14**, 66–75 (2012)
28. Elhamifar, E., Sapiro, G., Vidal, R.: Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In: *Advances in Neural Information Processing Systems*, pp. 19–27 (2012)
29. Elhamifar, E., Sapiro, G., Vidal, R.: See all by looking at a few: sparse modeling for finding representative objects. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1600–1607. IEEE Press, Providence (2012)
30. Das, A., Panda, R., Roy-Chowdhury, A.K.: Continuous adaptation of multi-camera person identification models through sparse non-redundant representative selection. *Comput. Vis. Image Underst.* **156**, 66–78 (2016)
31. Hadj Hassen, Y., Ayedi, W., Ouni, T., Jallouli, M.: Multi-shot person re-identification approach based key frame selection. In: *Proceedings of the Eighth International Conference on Machine Vision, International Society for Optics and Photonics, Barcelone*, p. 98751H (2015)
32. Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using spatial covariance regions of human body parts. In: *Proceedings of the Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 435–440. IEEE Press, Boston (2010)
33. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: *European Conference on Computer Vision*, pp. 688–703. Springer International Publishing, Zurich (2014)

Intelligent Interactive Multimedia Systems and Services
2017

De Pietro, G.; Gallo, L.; Howlett, R.J.; Jain, L.C. (Eds.)

2018, XV, 587 p. 224 illus., Hardcover

ISBN: 978-3-319-59479-8