Local Sparse Representation Based Interest Point Matching for Person Re-identification

Mohamed Ibn Khedher(✉) and Mounim A. El Yacoubi

Institut Mines-Telecom/Telecom SudParis: CEA Saclay Nano-Innov,
91191 Gif sur Yvette Cedex, France
{mohamed.ibn_khedher,mounim.el_yacoubi}@telecom-sudparis.eu

Abstract. This paper presents a multi-shot person re-identification system from video sequences based on Interest Points (SURFs) matching. Our objective is to improve the Interest Points (IPs) matching using low resolution images in terms of re-identification accuracy and running time. First, we propose a new method of SURF matching via Local Sparse Representation (LSR). Each SURF in the test video sequence is expressed as a sparse representation of a subset of SURFs in the reference dataset. Our approach consists of searching the latter subset from the reference IPs that are located on a similar spatial neighborhood to the query IP. Second, it investigates whether IPs filtering can decrease the re-identification running time. An ensemble of binary classifiers are evaluated. Our approach is assessed on the large dataset PRID-2011 and shown to outperform favorably with current state of the art.

Keywords: Person re-identification · Local sparse representation · Interest point · Filtering · Binary classifier · SURF

1 Introduction

Person re-identification is the task of determining if a person leaving the field of camera A reappears in the field of camera B. Re-identification may be viewed as a soft biometric task since it consists of matching a query input to a reference one by using low resolution information related to the human silhouette.

This work is an extension of the standard IPs matching via Sparse Representation (SR) [12] (It called later Standard SR). The idea behind is to express each query IP as a linear combination vector from a subset of reference IPs (called dictionary). The SR corresponds to a sparse vector whose nonzero entries correspond to the weights of reference IPs. To classify the query IP, a reconstruction error is calculated for each reference identity using only the SR coefficients corresponding to this identity. The query IP is then identified as the reference identity minimizing the reconstruction error. Finally, the reference person obtaining the majority of votes is claimed as the re-identified person.

With respect to standard SR, our contribution is twofold. First, the Standard SR uses all the reference dataset to construct the dictionary regardless of the

S. Arik et al. (Eds.): ICONIP 2015, Part III, LNCS 9491, pp. 241–250, 2015.
DOI: 10.1007/978-3-319-26555-1_28

spatial position of the query IP in the image. Therefore, due to IPs noisiness and ambiguity in uncontrolled conditions, IPs from different image parts may be included in the dictionary, thus making SR unreliable. In this work, we add a spatial constraint related to the position of IPs in the images. Concretely, we propose a Local Sparse Representation (LSR), where for each query IP, a dictionary is selected from its spatial neighborhood reference IPs. Second, rather than considering a binary 1/0 vote after SR matching, we propose to use a continuous vote for each identity, that is related to the weight of its nonzero coefficients in the SR.

The second part of this work is about IPs filtering. Having a large number of IPs per person, re-identification becomes a much time-consuming task. In this context, we propose a new filtering scheme to reject unreliable matched IPs pairs that are probably resulting from matching IPs from different persons or associated with different parts of the silhouette. To do this, we design a binary classifier that learns on a training dataset, made up of pairs of positive IP pairs (each pair {query IP, closest reference IP} is associated with the same person) and negative IP pairs (each pair {query IP, closest reference IP} is associated with different persons). Our motivation is that each IP for which the closest IP belongs to a different person is unreliable and is better to be dropped from SR matching and subsequent voting for re-identification. In this paper, we study the power of filtering of two popular classifiers, Support Vector Machine (SVM) and Random Forest (RF) and investigate the tradeoff between reducing running time and keeping a better re-identification performance.

The rest of the paper is organized as follows. In Sect. 2, a state of the art is presented. The principle of our approach is discussed in Sect. 3. Sections 4, 5 and 6 present respectively the major steps of the re-identification system: feature extraction, matching and filtering. Section 7 is dedicated to the experimental part and finally a conclusion and perspectives are presented.

2 State of the Art

From a learning perspective, the re-identification approaches can be grouped into two categories: supervised approach and unsupervised approach.

Unsupervised Approaches: This category mainly focuses on the way to represent the image. Usually, the latter is represented by a set of either IPs or regions corresponding generally to body parts. Hamdoun et al. [7] collect during a short video a set of SURFs to represent the person. The authors of [10] add a shape information to the standard SIFT to improve the matching step. In the category of region based approaches, Farenzena et al. [4] propose a Symmetry-Driven Accumulation of Local Features (SDALF) by exploiting the symmetry property of the human body and decomposing it into three parts. Authors of [9] combine color and texture features extracted from each rectangular region to form one vector descriptor per image.

Supervised Methods: The learning phase can be related to parameters of the metric used to compare images, or related to the discriminant descriptors selected

among all extracted features. Regarding learning metrics, in [6], an "Ensemble of Localized Features" (ELF) is presented to model person signature. The weights of features are learned using the Adaboost algorithm. In [8], the authors propose to learn a metric from pairs of samples from different cameras to take into account the transition between cameras. Authors of [1] propose to measure similarity between two images in a pre-learned space where correlation between images associated with the same person is maximized. As far as discriminative methods are concerned, a handful of works are found in the literature. Authors of [13] introduce a graph-based approach for a non-linear dimensionality reduction. It is applied to extract the most informative color representation to describe the person.

Our local sparse representation (LSR) method lies in the unsupervised category, while our filtering approach lies in the supervised one. LSR takes into account the spatial position of IPs in images contrary to [12]. Our filtering approach is automatic and does not depend on empirically parameters like in [11].

3 Proposed Approach

Our approach basically consists of four stages: (1) Feature extraction (SURFs), (2) SURFs Filtering based on binary classifier, (3) SURF identification via Local

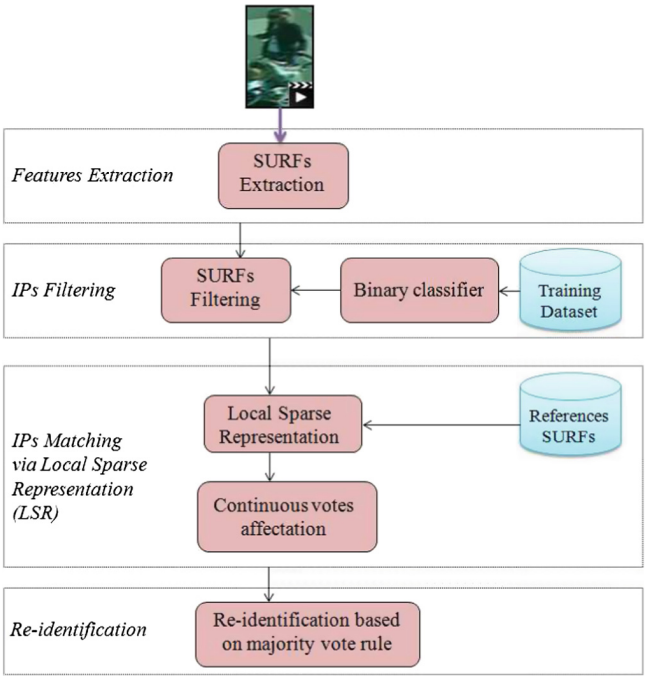


Fig. 1. Flowchart of our approach.

Sparse Representation (LSR) and (4) Person re-identification based on majority vote rule with continuous votes. Figure 1 shows the flowchart of our approach. Its principle is the following: first, each test person is described by a set of SURFs collected from a video sequence. A filtering step is then applied: after matching each test SURF to the closest reference one, we generate a difference vector, obtained by component-wise difference of the matched IP pair descriptors. The pre-learned classifier accepts or rejects the test IP based on the input difference. All retained SURFs are subsequently matched via LSR. To match one SURF, SR coefficients are used to infer a continuous vote for each reference identity based on its associated nonzero coefficients. In this way, a vote vector of dimension equal to the number of reference identities is generated. Finally, the reference person obtaining the majority of votes is claimed as the re-identified person.

4 Features Extraction

SURF is a popular IP descriptor proposed by [2], and used for several computer vision applications including person re-identification. We motivate our use of SURF by its robustness to geometric transformations (angle of view and scale) and to lighting variation and to its fast detection/description compared to others IPs. To compute SURF, two stages are required: SURF detection and SURF description. The detection step is based on the approximation of the determinant of Hessian matrix, while the descriptor is based on the Haar wavelet. The SURF descriptor considers a square region around the IP, divided into 4×4 grids to form 16 sub-regions. Four components related to Haar-wavelet x-responses and y-responses are extracted from each sub-region. Figure 2 shows samples of detected SURFs within an image from the used dataset.

5 SURF Matching via Local Sparse Representation (LSR)

Sparse representation consists of expressing a signal as a linear combination involving the smallest number of samples of a preselected dictionary. Given a query SURF q and a dictionnary A, SR finds the sparsest solution of the equation Eq. 1.

$$y = A\alpha \tag{1}$$

Our LSR is different from [15] in 2 points. First, in [15] a SR is calculated for the whole face, while ours is adapted to local features. Second, in [15], only one dictionary (the whole reference dataset) is used to compute SR for all test faces, while in our case we select a dynamic and reduced dictionary for each query SURF. From the other hand, our LSR is different from [12] in the way to select the dictionary. In [12], the dictionary is selected from all reference samples, while ours is selected from only the reference samples of the spatial neighborhood of the query IP.

The matching of one query SURF via LSR requires three steps: (1) Local dictionary selection, (2) Sparse representation and (3) Identity assignment.

- **Local Dictionary Selection:** the dictionary A is composed of the N closest reference SURFs. A is of dimension (DxN) where $D = 64$, dimension of SURF descriptor and N is empirically set to 200. The N closest reference SURFs are selected from the reference IPs in the spatial neighborhood, in a rectangular region around the query SURF, as shown in Fig. 2. The width of the region is learned on the training dataset; using this optimization, it is set in our experiments to 60 pixels. We use the same region dimensions when evaluating the unsupervised protocol.
- **Sparse Representation:** the Coordinate Descent Algorithm [5] is used to find the sparsest solution of Eq. 1 as shown in Eq. 2. Its advantage is the use of a tuning parameter λ, to adjust the tradeoff between sparsity term $\|\alpha\|_1$ and error reconstruction term $\|\Phi\alpha - y\|_2^2$.

$$\alpha_s = \min_{\alpha}(\|\Phi\alpha - y\|_2^2 + \lambda\|\alpha\|_1) \tag{2}$$

- **Identity Assignment:** the nonzero coefficients of α_s are used to identify the query IP. We propose to use a continuous vote contrary to [12] where a binary vote is generated. In fact, a x_i vector is calculated for each reference identity i having at least one non-zero coefficient:

$$x_i = [0, \ldots 0, \alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,k_i}, 0, \ldots, 0] \tag{3}$$

x_i is a coefficient vector obtained from α_s with all elements set to zero except those associated with the identity i. For each reference identity i, the associated vote V_i is incremented (Eq. 4) by a value reflecting the weight in the sparse representation of reference identity i.

$$V_i = V_i + \frac{\|x_i\|}{\|\alpha_s\|} \tag{4}$$

Then, the vote vector V is normalized to unit length. Finally, the query person is claimed as the person that gathers the majority of votes.

6 Binary Classifier for SURFs Filtering

The proposed filtering method is based on a supervised binary classifier. Its goal is to classify IPs into two classes: reliable and unreliable IPs. Ideally, the classifier discards unreliable IPs and retains reliable ones. Two classifiers are evaluated: Support Vector Machine (SVM) [14] and Random Forest (RF) [3]. To run a classifier, two stages are required: first, the classifier learns a filtering model which is used in the second step to discard or retain test IPs.

Training Stage: The classifier takes as input two vector sets: S_{Same} (positive vectors associated with class +1) and S_{Diff} (negative vectors associated with

Fig. 2. Local Dictionary Selection: Left (samples of test SURFs), Right (reference dataset). To match a query SURF (green point), a dictionary is selected from all reference SURFs belongs to a rectangular regions around the test SURF (Points in the violet region) (Colour figure online).

class -1). S_{Same} and S_{Diff} model respectively reliable and unreliable IPs. To construct S_{Same} and S_{Diff}, each query SURF is matched to its closet reference one; if the matched pair is associated with the same person, the difference pair descriptor is added to S_{Same}, else it is added to S_{Diff}. In the case of SVM, the training stage consists of finding the hyperplane that separates S_{Same} and S_{Diff}; while for RF, these two sets are used to construct trees by maximizing the variance between the two classes.

Test Stage: To classify a query IP, SVM uses the pre-learned model to assign a probability to each class. The IP is retained if $\mathrm{P}(+1) > \mathrm{P}(-1)$, where P(.) is a function returning the probability of input class. On the other hand, RF classifies a query IP by running down all of the tree. Then tree decisions (predicted classes) are aggregated to provide a final decision (majority vote rule).

7 Experimental Results

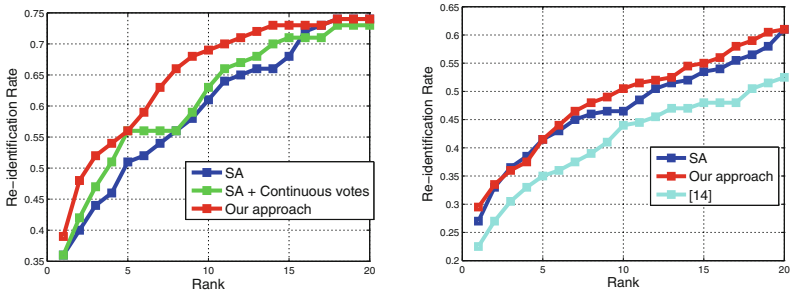
We evaluated our approach on the multi-shot dataset PRID-2011 obtained from two cameras (A and B). The camera-A filmed 749 people and Camera-B filmed 385 people (200 people are common). Two protocols are used in evaluation:

- Unsupervised Protocol consists of identifying the 200 common people filmed by Camera-A in the gallery set (Camera-B) of 749 people.
- Supervised Protocol: PRID-2011 is divided into two parts: training and test. The training set contains two sequences of the first 100 common people. The test set contains the remainder 649 people from Camera-B in reference and the remaining common 100 people from Camera-A in test.

Results are shown in terms of the Cumulative Matching Characteristic (CMC) curve associated with the identification rate. Throughout the rest of this paper, Standard Approach (SA) means that (1) the dictionary is selected from all reference SURFs, (2) binary votes are used and (3) non filtering is applied.

7.1 Contribution of LSR with Continuous Votes

We evaluated our LSR on the PRID-2011 dataset using the two protocols. Starting by the supervised one, results are shown in Fig. 3a and Table 1. Figure 3a shows the obtained CMC (From rank 1 to rank 20) of our approach compared to SA and (SA + continuous votes). Table 1 shows different methods performances (identification rate at rank 1).



(a) Supervised protocol (b) Unsupervised protocol

Fig. 3. CMC performance on PRID-2011

Table 1. Results on PRID-2011 (Supervised protocol)

Approach	Re-identification rate (%)
Standard Approach (SA)	36
SA + continuous votes	36
Our approach	39

The PRID-2011 is evaluated in the state of the art only using the unsupervised protocol where all the dataset is used in test. The obtained results with the unsupervised protocol are shown in Fig. 3b and Table 2. Figure 3b shows the obtained CMC (From rank 1 to rank 20) compared to SA and the state of the art. Table 2 shows our performance compared to the state of the art.

For both protocols, the results show that our approach outperforms the standard one (SA). Using the supervised protocol, our approach achieves an improvement of 3 % in the re-identification rate at rank 1. This proves that adding a spatial constraint to construct the dictionary makes sparse representation more effective. Moreover, it proves the efficiency of using continuous votes (soft decisions) rather than binary votes (hard decisions). Using the unsupervised protocol, the results show the benefits of (LSR + continuous votes). Our approach achieves an improvement of 2.5 % in the re-identification rate at rank 1

Table 2. Results on PRID-2011 (Unsupervised protocol)

Approach	Re-identification rate (%)
[9]	19.18
[11]	22.5
[12]	27
Our approach	29.5

w.r.t SA. Compared to the state of the art, our approach achieves an improvement of 12.32 % in the e-identification rate when compared to [9] and 7 % when compared to [11]. This improvement is very significant given the large size of the dataset.

7.2 Contribution of IPs Filtering

We evaluated our IPs' filtering method on PRID-2011 using the supervised protocol. Table 3 compares the results obtained after IPs filtering using one of the two classifiers (SVM or RF) with those of the system where no filtering is applied, according to two performance indicators: re-identification rate and average running time per image.

Table 3. Results of our approach on PRID-2011 (Supervised Protocol)

Classifier	Filtering rate	Re-identification rate	Running time/Image
RF	56.81 %	38 %	1.31(s)
SVM	78.93 %	39 %	0.92(s)
—	No filtering	39 %	2.36(s)

Table 3 shows that by filtering 56.81 % of IPs using RF or 78.93 % using SVM, the accuracy of our approach does not decrease while the processing time becomes much lower. For example, the IPs filtering of SVM achieves an improvement of 61.01 % in average running time per image. These results prove the importance of filtering to reduce running time.

8 Conclusion

This paper has studied IPs matching in uncontrolled conditions for a human re-identification task. It proposed a novel IP matching via Local Spare Representation (LSR). The idea behind is to take into account the spatial distribution of IPs in reference and test images. Our contribution consists of selecting the dictionary from only the reference IPs lying on a learned spatial neighborhood

of the query IP. Moreover, we used a soft IPs identification based on continuous votes. On the other hand, we proved the importance of IP filtering to reduce the re-identification running time. The experiment results on the large PRID-2011 database showed that our LSR method performed better in terms of re-identification rate when compared to the state of the art. Moreover it proved the utility of IPs filtering to reduce running time. Using SVM for IPs filtering allows to automatically discard about 80 % of the IPs in the test dataset while keeping the same re-identification accuracy when processing all IPs without filtering. In the future, we will focus on optimizing the Local Dictionary since the latter's size affects significantly the total running time. Moreover, we will study better SR representation schemes for the re-identification task.

References

1. An, L., Kafai, M., Yang, S., Bhanu, B.: Reference-based person re-identification. In: Proceedings of the 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 244–249 (2013)
2. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
4. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: Conference on Computer Vision and Pattern Recognition, pp. 2360–2367 (2010)
5. Friedman, J.H., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**, 1–22 (2010)
6. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 262–275. Springer, Heidelberg (2008)
7. Hamdoun, O.: Pedestrian detection and re-identification using interest points between non overlapping cameras. Ph.D. thesis, École Nationale Supérieure des Mines de Paris (2010)
8. Hirzer, M., Roth, P., Bischof, H.: Person re-identification by efficient impostor-based metric learning. In: Proceedings of the 9th IEEE International Conference on Advanced Video and Signal-Based Surveillance, pp. 203–208 (2012)
9. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 91–102. Springer, Heidelberg (2011)
10. Jungling, K., Arens, M.: View-invariant person re-identification with an implicit shape model. In: International Conference on Advanced Video and Signal-Based Surveillance, pp. 197–202 (2011)
11. Khedher, M.I., El-Yacoubi, M.A., Dorizzi, B.: Probabilistic matching pair selection for surf-based person re-identification. In: International Conference of Biometrics Special Interest Group, pp. 1–6 (2012)
12. Khedher, M.I., El-Yacoubi, M.A., Dorizzi, B.: Multi-shot surf-based person re-identification via sparse representation. In: International Conference on Advanced Video and Signal-Based Surveillance (2013)

13. Cong, D.N.T., Achard, C., Khoudour, L., Douadi, L.: Video sequences association for people re-identification across multiple non-overlapping cameras. In: Foggia, P., Sansone, C., Vento, M. (eds.) ICIAP 2009. LNCS, vol. 5716. Springer, Heidelberg (2009)
14. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience, New York (1998)
15. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. **31**, 210–227 (2009)