Multiple Scaled Person Re-Identification Framework for HD Video Surveillance Application

Hua Yang[1,2](✉), Xinyu Wang[1,2], Wenqi Ma[1,2], Hang Su[1,2], and Ji Zhu[1,2]

[1] Institue of Image Communication and Network Engineering,
Shanghai Jiao Tong University, Shanghai 200240, China
[2] Shanghai Key Lab of Digital Media Processing and Transmission,
Shanghai 200240, China
{hyang,xinyuwang}@sjtu.edu.cn, mawenqi@sjtu.org,
{suhangss,jizhu1023}@gmail.com

Abstract. Person re-identification is an important problem in automated video surveillance. It remains challenging in terms of extraction of reliable and distinctive features, and matching of the features under different camera views. In this paper, we propose a novel re-identification strategy for person re-identification based on multiple image scaled framework. Specifically, global features and local features are extracted separately in different image scales. These two-scaled processing are constructed in a cascaded system. We use semi-supervised SVM to obtain a similarity function for global features and a similarity function combining the spatial constraint and salience weight for local features. Experiments are conducted on two datasets: ETHZ and our dataset with high resolution. Experimental results demonstrate that the proposed method outperforms the conventional method in terms of both accuracy and efficiency.

Keywords: Person re-identification · Multiple scaled framework · Distance metrics

1 Introduction

Person re-identification across different views of cameras is a fundamental task in automated video surveillance.Despite best efforts have been made in computer vision area in the past years, person re-identification problem remains largely unsolved. This is due to a number of reasons. First, the resolution of the current monitored cameras is not high enough so that person verification relying upon biometrics is infeasible and unreliable. Second, as the transition time between disjoint cameras varies greatly from individual to individual with uncertainty, it is hard to impose accurate temporal and spatial constraints. Third, the visual appearance features, which are extracted mainly from the clothing and shapes of people, are intrinsically indistinctive for matching people.

To solve the re-identification problem, discriminative and reliable signature for the person is needed. The image can be described by color[1], shape[2, 3],

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 219–228, 2015.
DOI: 10.1007/978-3-662-48570-5_21

texture[3, 4, 5, 6], Haar-like representations[7], edges[3], interest points[8, 9, 10] and image patches[4]. Since a single type of features is not powerful enough to capture the subtle differences of all pairs of objects, multiple features are combined here to make the person signatures more discriminative and reliable. Bazzani et al.[1] and Cheng et al. [11]combined MSCR descriptors with weighted Color Histograms, achieving state-of-the-art results on several widely used person re-identification datasets.There are also some other research works on person re-identification have been done to learn reliable and effective mid-level features. Li et al. [12] proposed a deep learning framework to learn filter pairs, which encode photometric transforms across camera views for person re-identification. Zhao et al. [13] proposed a method to automatically learn discriminative mid-level features without annotation of human attributes.

Conventional methods attend to seek the discriminative and reliable signature, after feature extraction, these methods simply choose a standard distance measure such as L1 norm and L2 norm. However, under severe changes in viewing conditions, extracting a set of features that are both distinctive and reliable is extremely hard. Moreover, given that certain features could be more reliable than others under a certain condition, applying a standard distance measure is undesirable as it essentially treats all features equally without discarding bad features selectively in certain matching circumstances. To overcome these shortcomings, recent years researchers focus on the distance metric of person re-identification. That is, given a set of features of each person image, they seek to quantify and differentiate these features by learning the optimal distance measure that is most likely to give correct matches. Gray et al. [4] combined spatial and color information in an ensemble of local features by boosting. Prosser et al. [5] formulated person re-identification as a ranking problem, and used ensembled RankSVMs to learn pairwise similarity.Zheng et al. [14] formulate person re-identification as a relative distance comparison learning problem in order to learn the optimal similarity measure between a pair of person images.

This paper focuses on effectively using the benefits of high resolution images and reducing the complexity in the foundation of improving the accuracy of re-identification. To perform re-identification under the HD monitor cameras, we propose a re-identification framework, in which global features and local features are extracted respectively in different image scales based on their scale behaviours. Specifically, the global features are represented by the histogram whose matching performance is not related to the image scale directly while the local features perform better on the higher image scale since they need more feature details to match. It is worth mentioning that our approach is performed under special application scenarios, that is the scenarios monitored by HD cameras. So we do not compare with the most advanced methods.

The contributions of this framework can be summarized in three-folds: First, we propose a novel multiple scaled framework in which different features can be extracted in proper image scales on their behaviors, so that the benefits of HD monitor cameras can be exploited. Second, we use a cascaded system to improve the efficiency of the system. Third, we use a new matching algorithm based on feature points, adding the color feature into the descriptor and combing the spatial constraint and salience weight, which improves the accuracy of person re-identification.

The rest of the paper is organized as follows: Section 2 describes the details of the proposed framework which we refer to as Multiple Scale Re-identification Framework (MSRF). Section 3 illustrates and analyzes the experimental results. Finally, the main conclusions are summarized in Section 4.

2 Multiple Scale Re-identification Framework

Figure 1 shows a re-identification system under the HD monitor circumstances. There are four steps in our framework. First, the global features are extracted on low image scale. Second, we use semi-supervised SVM to get a match result and choose the top $k\%$ as the filtered candidate. Third, local features are extracted on high image scale. Forth, a local feature points based algorithm is used to get another match result. Last,the match results obtained in the two-scaled processing are added together to get the final ranking. In our experiments, the low image scale is obtained with the down sampling factor 2 while the large scale with the sampling factor 1.

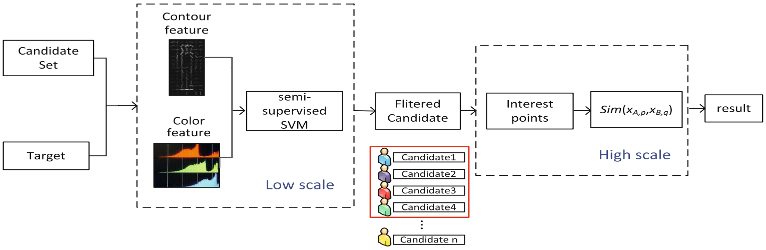


Fig. 1. System structure of Multiple Scale Re-identification

2.1 Re-identification on Low Image Scale

The appearance of objects is usually characterized in three aspects, color, contour and texture. Since a single type of feature is not powerful enough to capture the subtle differences of all pairs of objects, color and contour are combined here to make the person signatures more discriminative and reliable.

Color. Color histograms of the whole image region are widely used as global features to match objects across camera views because they are robust to the variations of poses and viewpoints. However, they also have the weakness that they are sensitive to the variations of lighting conditions and photometric settings of cameras and that their discriminative power is not high enough to distinguish a large number of objects. Various color spaces such RGB, Lab, HSV and Log-RGB have been investigated and compared in [15]. Removing the lightness component in the HSV color space can reduce the color variation across camera views and we use this method to obtain color feature in our experiment.

Contour. Histogram of Oriented Gradients (HOG)[15]characterizes local shapes by capturing edges and gradient structures. It is robust to small translations and rotations of object parts.

In our experiments, the color feature and contour feature are concatenated to form a new representation.The dimension of the histogram in each color channel is 128 and the dimension of HOG histogram is 3528.

After the feature histograms have been extracted, the standard distance measure such as L1 norm could be applied. The distance learning method like RankSVM [5], RDC [14] could also be used according to the application circumstances. Here we use semi-supervised SVM to get a match result and choose the top $k\%$ as the filtered candidate.Here is a brief introduction of the principle of semi-supervised SVM. In order to exploit the unlabeled data, Bennett[16] created a method to classify the unlabeled data based on the original support vector machine. It is assumed that the unlabeled points are classified as Category 1, and the classification accuracy is calculated. Then, the points are classified as Category 2. Select the class that has the high classification accuracy.

The choice of k relates to the accuracy and complexity of the algorithm. If k is too small,the number of samples for the following processing is small,which will reduce the accuracy of the algorithm.If k is too large, the complexity will be increased. Considering the accuracy and complexity, we choose k=30 here.

2.2 Re-identification on High Image Scale

Local features perform better on high image scale since they need more feature details to match.Under the HD monitor circumstances, the image details could be exploited, which will benefit the matching performance based on the texture interest points. Traditional methods just extracted the texture feature of the interest points, which were less discriminative and reliable. Here this paper proposed an improved re-identification method based on the interest points, which we called Local Salience Feature (LSF) method.

There are four steps in the re-identification. First, the interest points will be extracted by the SURF operator. Second, in the center of each interest point, a patch is extracted. Then the color histogram will be extracted from the patch. Color histogram and contour histogram are concatenated to form a new representation. Third, the location of each point is considered to make the matching process more effective. Finally, the salience weight of each point is learned to make the re-identification more reliable.

Feature Extraction. In the center of each interest point, a patch will be extracted. A LAB color histogram is extracted from each patch. For the purpose of combination with other features, all the histograms are L2 normalized. To handle viewpoint and illumination changes, SURF descriptor[17]is used as a complementary feature to color histograms, which are also L2 normalized. In our experiments, the parameters of feature extraction are as follows: patches of size 10x10 pixels.128-bin color histograms are computed in L, A, B channels respectively. And in each channel, SURF features produces a 128 feature vector for each interest points. In a summary, each patch is finally represented by a discriminative descriptor vector with length 128x3+128 =512.

Spatial Constrain. The distance of pairwise person could be converted to compute the distance of pairwise interest point set. The greedy algorithm [18] will be applied in our experiment. For each point of the target, we will find the corresponding one which has the shortest distance with it in the candidate point set. For each point pair, in order to deal with the misalignment in the matching process, we also compute the distance of the locations of the features with the Euclidian distance. The final ground distance between two interest points is shown in Eq.1.

$$D\left(x_A, x_B\right) = FD\left(x_A, x_B\right) + \alpha \times ED\left(x_A, x_B\right) \tag{1}$$

Where α is a weighting parameter, FD is the distance of the feature vector, L1 norm is applied in our experiments, ED is the Euclidean distance, and x is the location of the centroid of the point. It is worth mentioning that several distance measures were considered, and experiments showed the effectiveness of the proposed combination of distances.

As suggested in [15], aggregating similarity scores is much more effective than minimizing accumulated distances, especially for those misaligned or background points which could generate very large distances during matching. By converting to similarity, their effect could be reduced. We convert distance value to similarity score with the Gaussian function:

$$s\left(x_A, x_B\right) = exp\left(-\frac{D\left(x_A, x_B\right)^2}{2}\right) \tag{2}$$

Salience Weight. Each interest point of a person has certain information, so different point has different identify power in the matching process. According to [19], KNN method could be applied to learning the salience weight of the interest points. We could get the following function:

$$X_{nn}\left(x^i_{A,p}\right) = \left\{ x | arg \max_{x^j_{B,q} \in \{B_q\}} s\left(x^i_{A,p}, x^j_{B,q}\right), q = 1, 2, ..., N_B \right\} \tag{3}$$

Where the interest point in target image is represented as $x^i_{A,p}$, where (A, p) denotes the p-th image in camera A and i denotes the point index in set. The interest point in candidate image is represented as $x^j_{B,q}$, where (B, q) denotes the q-th image in camera B and j denotes the point index. $\{B_q\}$ means the candidate set under camera B. s is the similarity score function in Eq.2. N_B is the candidate number.

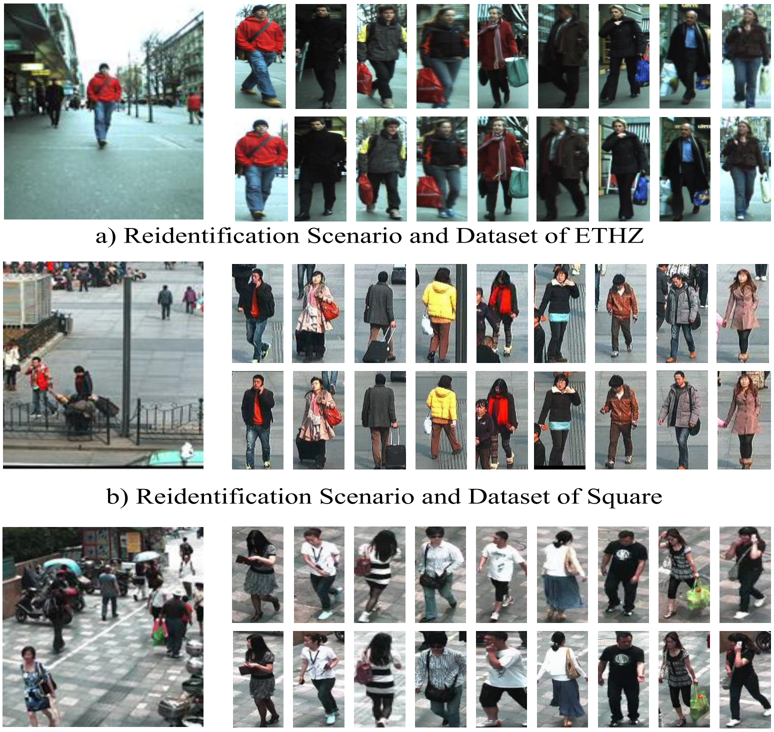
We apply a similar scheme in [19] for each test point, and the KNN distance is utilized to define the salience score:

$$w\left(x^i_{A,p}\right) = D\left(x^i_{A,p}, X_{nn}\left(x^i_{A,p}, k\right)\right), k = \frac{N_B}{2} \tag{4}$$

Where D denotes the distance of the k-th nearest neighbor. If the distribution of the reference set well reflects the test scenario, the interest point could only find limited number of visually similar neighbors. More details about the salience learning method could be found in [19][2].

Then we could get the similarity of two point sets by the following equation.

$$sim\left(x_{A,p}, x_{B,q}\right) = \sum_{i=1}^{|x_{A,p}|} w\left(x^i_{A,p}\right) \cdot s\left(x^i_{A,p}, x^j_{B,q}\right) \tag{5}$$



a) Reidentification Scenario and Dataset of ETHZ

b) Reidentification Scenario and Dataset of Square

c) Reidentification Scenario and Dataset of Road

Fig. 2. Example images of different datasets used in our evaluation.The first column denotes the scenario and the rest columns denote image pairs of the same person.**a)** ETHZ,**b)** our dataset on square,**c)** our dataset on road

3 Experimental Results

3.1 Experiment 1

In this experiment, we evaluated the accuracy of the proposed strategy. We evaluated our approach on the public ETHZ dataset. The results are shown in standard Cumulated Matching Characteristics (CMC) curve [19]. A rank r matching rate indicates the percentage of the p images with correct matches found in the top r ranks against the p gallery images. Rank 1 matching rate is thus the correct matching/recognition rate. Note that, in practice, although a high rank 1 matching rate is critical, the top r ranked matching rate with a small r value is also important because that the top matched images will normally be verified by a human operator. We also apply our method on two real HD monitor circumstances. Both ETHZ and real dataset are shown in Fig. 2.

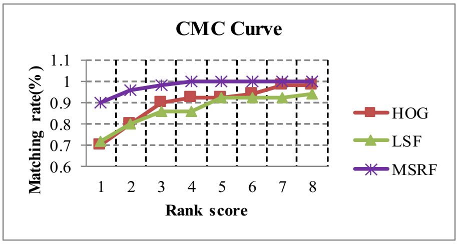


Fig. 3. Re-identification Result of ETHZ Dataset

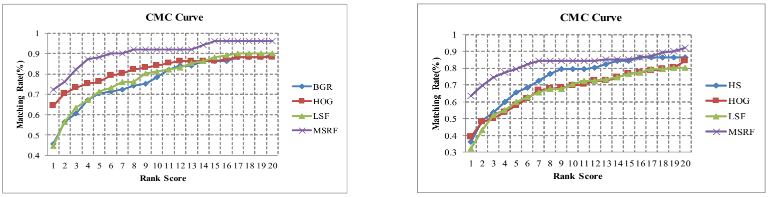
Table 1. Matching rate(%) of Different Methods on ETHZ Dataset

	ETHZ Dataset				
Methods	r=1	r=2	r=3	r=4	r=5
MSRF	90.00	96.00	98.00	100.00	100.00
BGR[15]	62.00	76.00	82.00	88.00	92.00
HS[15]	62.00	88.00	92.00	96.00	100.00
LAB[15]	66.00	86.00	92.00	98.00	100.00
HOG[18]	70.00	80.00	90.00	92.00	92.00
SIFT[18]	46.00	56.00	58.00	62.00	64.00
LSF	72.00	80.00	86.00	86.00	92.00

ETHZ Dataset. This dataset contains video sequences captured from moving cameras. It contains a large number of different people in uncontrolled conditions. With these video sequences, we get 50 pairwise people images for evaluation. All image samples are normalized to 128x64 pixels. Traditional methods like appearance-based methods[15]are compared here.

As shown in Fig. 3 and Table 1, our approach outperforms other methods based on single feature because that the MSRF method exploits the benefits of the HD images and multiple features are combined in our framework. The ETHZ is not a very challenging dataset,so we evaluate our method on two real monitor circumstances with different challenges.

Square and Road Dataset. The square datasets were captured from a railway station by two non-overlapping cameras. We collected 101 pairwise people images under each monitor circumstance for evaluation. All image samples are normalized to 128x64 pixels. Traditional methods like appearance-based methods[15]are compared here.



a) Scenario of Square

b) Scenario of Road

Fig. 4. Re-identification Result of Real Monitor Dataset

Table 2. Matching rate (%) of Different Methods on Different Datasets

	Square scenario				*Road scenario*			
Methods	r=1	r=5	r=10	r=20	r=1	r=5	r=10	r=20
MSRF	72.28	88.12	92.79	96.04	63.72	77.45	84.31	92.15
BGR[15]	45.54	70.30	78.22	89.11	22.55	47.06	66.67	78.41
HS[15]	44.55	63.37	74.26	83.17	36.27	59.80	79.41	86.27
LAB[15]	44.55	68.32	78.21	89.11	21.57	50.00	67.64	79.41
HOG[18]	64.36	76.24	84.16	88.12	39.21	53.92	69.61	84.31
SIFT[18]	55.45	65.35	73.27	82.18	24.72	36.49	50.21	71.78
SURF[18]	56.44	70.30	79.21	83.17	24.90	40.69	74.51	75.00
LSF	57.56	71.28	81.19	90.10	32.35	54.90	70.59	80.39

As shown in Fig. 4 and Table 2, our approach outperforms other methods based on single feature, especially when r is small. This is because the MSRF method exploits the benefits of the HD images and multiple features are combined in our framework. Images have a range of variations in human appearance and illumination under the real monitor circumstances, which made single feature less discriminative and reliable. In our framework, multiple features will have more identify power which benefited the re-identification result. Moreover, the local feature extracted on the high scale can get more feature details, which made the match result more reliable.

3.2 Experiment 2

In this experiment, we evaluated the high efficiency of the proposed strategy. Table 3 gives the cost of different algorithms. The hardware platform is Intel i7, 3.4GHz, 4GB RAM. Each algorithm is conducted on 101 pairwise people images and there are 10210 comparison operations in our experiments. It can be seen that the algorithms based on statistical characteristics have low complexity because the number of distance calculations is small. While the local feature algorithms need to conduct matching operation for each feature point. For there are M feature points extracted from one pedestrian image and N feature points from another pedestrian image, the computation cost is *O(MN)*. In the proposed strategy, firstly we use the statistical characteristics based algorithm to obtain the selected candidates. And then we use the local feature points based algorithm to recognise the selected candidates. These two steps can reduce the complexity.

Table 3. Time Cost Result of Different Methods

Methods	COLOR	CONTOUR	SURF	LSF	MSRF
Cost Time(ms)	14992	71480	842046	842311	307806

4 Conclusions

In this paper, we propose a new re-identification framework based on the multiple scaled framework to perform re-identification under the HD monitor cameras. Global features and local features are extracted separately in different image scales based on their scale behaviours. Specifically, the global features are represented by the histogram whose matching performance is not related to the image scale directly, while the local features perform better on the higher image scale since they need more feature details to match. In our framework, firstly we use the statistical characteristics based algorithm to obtain the selected candidates. And then we use the local feature points based algorithm to recognise the selected candidates. Experimental results demonstrate that the proposed method outperforms the conventional method in terms of re-identification accuracy and efficiency.

Acknowledgments. This research is supported by STCSM (No.10231204002, No.11231203102),NSFC (No.61171172) and the Innovation Research Group Project of National Natural Science Foundation of China(No.61221001).

References

1. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. Comput. Vis. Image Underst. **117**(2), 130–144 (2013)

2. Oreifej, O., Mehran, R., Shah, M.: Human identity recognition in aerial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2010)
3. Schwartz, W., Davis, L.: Learning discriminative appearance based models using partial least squares. In: Brazilian Symposium on Computer Graphics and Image Processing (2009)
4. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European Conference on Computer-Vision, pp. 262–275 (2008)
5. Prosser, B., Zheng, W., Gong, S., Xiang, T.: Person re-identification by support vector ranking. In: Proceedings of the British Machine Vision Conference (2010)
6. Zhang, Y., Li, S.: Gabor-LBP based region covariance descriptor for person re-identification. In: International Conference on Image and Graphics, pp. 368–371 (2011)
7. Bak, S., Corvee, E., Bremond, F., Thonnat, M.: Person re-identification using haar-based and DCD-based signature. In: Proceedings of International Workshop on Activity Monitoring by Multi-camera Surveillance Systems (2010)
8. Gheissari, N., Sebastian, T., Tu, P., Rittscher, J., Hartley, R.: Person reidentification using spatiotemporal appearance. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition **2**, 1528–1535 (2006)
9. Kai, J., Bodensteiner, C., Arens, M.: Person re-identification in multi-camera networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio Workshops, pp. 55–61 (2011)
10. Zheng, W., Gong, S., Xiang, T.: Associating groups of people. In: Proceedings of British Machine Vision Conference (2009)
11. Cheng, D., Cristani, M., Stoppa, M., Bazzani, L., Murino, V.: Custom pictorial structures for re-identification. In: Proceedings of British Machine Vision Conference (2011)
12. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
13. Zhao, R., Ouyang, W., Wang, X.: Learning midlevel filters for person reidentification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
14. Zheng, W.S., Gong, S., Xiang, T.: Reidentification by relative distance comparison. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 653–668 (2013)
15. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: Proceedings of the IEEE International Conference on Computer Vision (2007)
16. Bennett, K., Demiriz, A.: Semi-supervised support vector machines. In Advances in Neural Information Processing Systems 11 (1998)
17. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Proceedings of the European Conference on Computer Vision, pp. 404–417 (2006)
18. Doretto, G., Sebastian, T., Tu, P.H., Rittscher, J.: Appearance-based person reidentification in camera networks: problem overview and current approaches. J. Ambient Intell. HumanizedComput. **2**(2), 127–151 (2011)
19. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2013)
20. http://www.vision.ee.ethz.ch/aess/dataset/