PERSON RE-IDENTIFICATION IN MULTI-CAMERA SYSTEM BY SIGNATURE BASED ON INTEREST POINT DESCRIPTORS COLLECTED ON SHORT VIDEO SEQUENCES

Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu and Bruno Steux

Robotics laboratory (CAOR)
Mines ParisTech
60 Bd St Michel, F-75006 Paris, FRANCE

ABSTRACT

We present and evaluate a person re-identification scheme for multi-camera surveillance system. Our approach uses matching of signatures based on interest-points descriptors collected on short video sequences. One of the originalities of our method is to accumulate interest points on several sufficiently time-spaced images during person tracking within each camera, in order to capture appearance variability.
A first experimental evaluation conducted on a publicly available set of low-resolution videos in a commercial mall shows very promising inter-camera person re-identification performances (a precision of 82% for a recall of 78%).
It should also be noted that our matching method is very fast: ~ 1/8s for re-identification of one target person among 10 previously seen persons, and a logarithmic dependence with the number of stored person models, making re-identification among hundreds of persons *computationally* feasible in less than ~ 1/5 second.

Index Terms— Video-surveillance, camera networks, person identification and tracking, re-identification, interest points

1. INTRODUCTION AND RELATED WORKS

In many video-surveillance applications, it is desirable to determine if a presently visible person has already been observed somewhere else in the network of cameras. This kind of problematic is commonly known as "person re-identification", and a general presentation of this field can be found for instance in §7 of [1]. Re-identification algorithms have to be robust even in challenging situations caused by differences in camera viewpoints and orientations, varying lighting conditions, pose variability of persons, and rapid change in clothes appearance.

A first category of person re-identification methods rely on biometric techniques (such as face or gait recognition), but in this paper we focus on the second group of methods, which use only global appearance. Among these, various approaches have been proposed: signature based on color histograms (such as in [2] or [3]), texture characteristics (see eg [4]), or panoramic model from multi-view [11]. A more extensive survey of various kinds of object signatures can be found in [12]. More recently some works have proposed the use of matching of interest points for establishing correspondance between objects, like cars in [5], and also for person re-identification as for instance in [6].

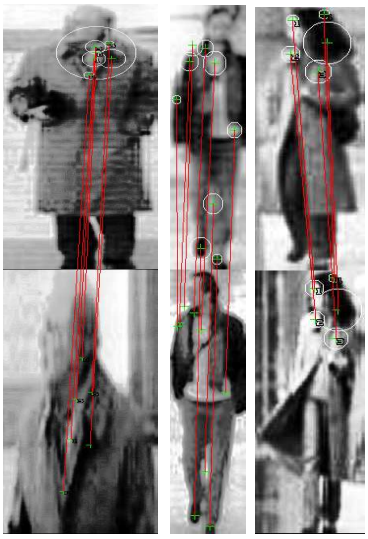

Figure 1: examples of successful matching of interest points for a given person seen under 2 different viewpoints and scales(left), with pose change (middle), and with clothe appearance and illumination change (right).

We here propose and evaluate a person re-identification scheme using matching of interest points collected in several images during short video sequences. The central point of our algorithm lies in the exploitation of image sequences,

978-1-4244-2665-2/08/$25.00 ©2008 IEEE

contrary to the method proposed in [6] where matches are done on image-to-image basis. This allows to get a more "dynamic" and multi-view descriptor than when using single image, and is a bit similar in spirit to the "averaging of interest-point descriptors throughout time sequence" used in the work by Sivic and Zisserman in [7], or the use (for content-based video matching) of interest point descriptors found in "video volumes" in [13]. However, contrary to both of them, we do not use SIFT [8] detector and descriptor, but a locally-developped (see §2) and particularly efficient variant of SURF [9]. This is also in contrast with Gheissari et al. in [6] who use a color-histogram of the region around interest points for their matching. Note also that we do not use a "vocabulary" approach as in [5] or [7], nor a "cluster" representation of descriptors as in [13], but perform a direct very efficient matching between interest point descriptors.

2. DESCRIPTION OF OUR PERSON RE-IDENTIFICATION SCHEME

In this section we detail the algorithmic choices made in our person re-identification approach. Our method follows the classical scheme of Detection-Recognition-Identification (DRI) algorithms. It can be separated in two main stages: a learning phase, and a recognition phase, as illustrated on figure 2a and 2b. The learning consists in taking advantage of tracking of a given person in a sequence from one camera, in order to extract interest points and their descriptors necessary to build the model consisting of a signature for each person.

The interest point detection and descriptor computation is done using "key-points" functions available in the Camellia image processing library developed in our lab (http://camellia.sourceforge.net). These Camellia key-points detection and characterization functions, which shall be described elsewhere in more details, are implementing a very quick variant inspired from SURF [9] but even faster. SURF itself is a recently proposed and extremely efficient alternative to the more classic and widely used interest point detector and descriptor SIFT [8]. In Camellia "key-points" as in SURF, the detection of points is Hessian-based (rather than Laplacian-based) and uses integral image for a very efficient computation. The point descriptors, whose computation also takes advantage of integral image, are 64D vectors coarsely describing distribution of Haar-wavelet responses in 4x4 sub-regions around the interest point (see [9] for more details). Camellia key-points are an optimized implementation of a variant of SURF, using integer-only computations, which makes it particularly well-suited for embedding in camera hardware.

Our recognition step uses tracking of the to-be-identified-person, and models built during learning stage, in order to determine if the signature of the currently analyzed person is similar enough to one of those already "registered" for which signatures have been stored from other cameras. Our method can be detailed in the following 5 steps:

1. **Model building**
   A signature is built for each detected and tracked person. In order to maximize the quantity of non-redundant information, we do not use every successive frame, but instead images sampled every half-second. The person's signature is obtained as the accumulation of interest point descriptors obtained on those images.
2. **Query building**
   The query for the target persons is built on several evenly time-spaced images, exactly in the same way as the models, but with a smaller number of images (therefore collected in a shorter time interval).
3. **Descriptor comparison**
   The metric used for measuring the similarity of interest point descriptors is the Sum of Absolute Differences (SAD).
4. **Robust fast matching**
   A robust and very fast matching between descriptors is done by the employed Camellia function, which implements a Best Bin First (BBF) search in a KD-tree [10] containing all models.
5. **Identification**
   The association of the query to one of the models is done by a voting approach: every interest point extracted from the query is compared to all models points stored in the KD-tree, and a vote is added for each model containing a close enough descriptor; finally the identification is made with the highest voted-for model.

3. EXPERIMENTAL EVALUATION

Direct comparison with other existing approaches is not easy, as to our knowledge there is no available benchmark for evaluation of person re-identification. We therefore decided to conduct a first experimental evaluation of our proposed method on a *publicly available* series of videos (http://homepages.inf.ed.ac.uk/rbf/CAVIAR) showing persons recorded in corridors of a commercial mall. These videos (collected in the context of European project CAVIAR [IST 2001 37540]) are of relatively low resolution, and include images of the same 10 persons seen by two cameras with very different viewpoints (see figure 3).

The model for each person was built with 21 evenly time-spaced images (separated by half-second interval), and each query was built with 6 images representing a 3 second video sequence (see figure 4). Camera color potential variability is avoided by working in grayscale. Illumination invariance is ensured by histogram equalization of each person's bounding-box.

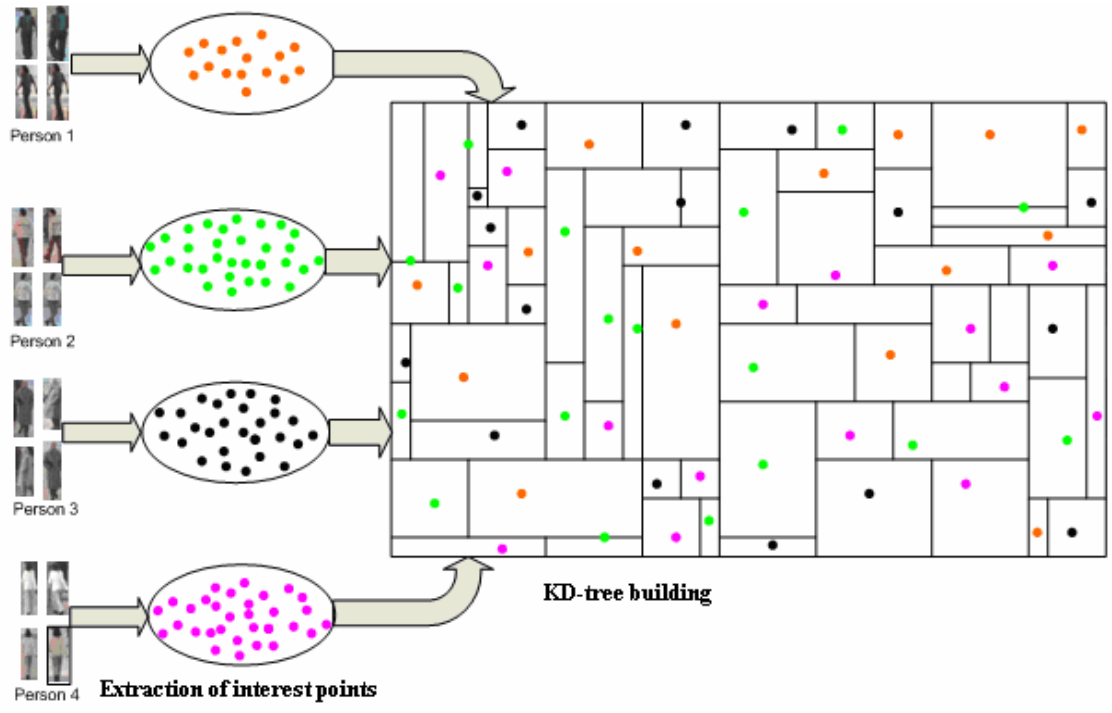

Figure 2a: Schematic view of model building.
For each tracked person, interest points are collected every 10 frames, and the person's signature is the union of all these key-points stored in a global KD-tree.

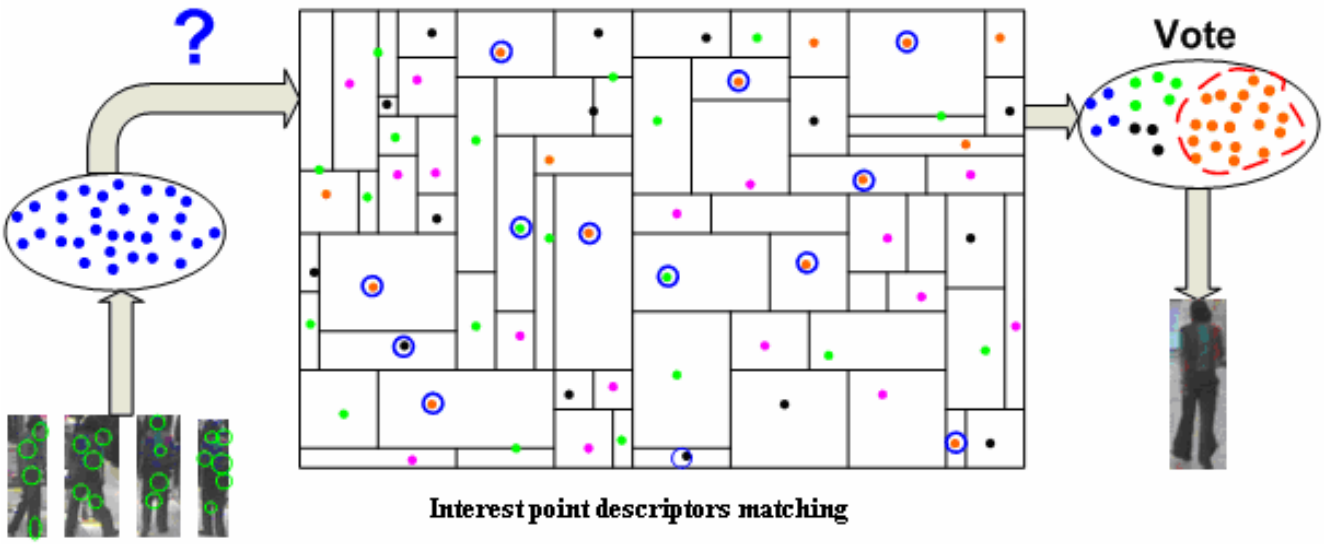

Figure 2b: Schematic view of re-identification of a query.
Interest points are collected similarly during tracking of a person to be re-identified, appearing in another camera, and a vote is made according to the respective identifications associated to all matching key-points.

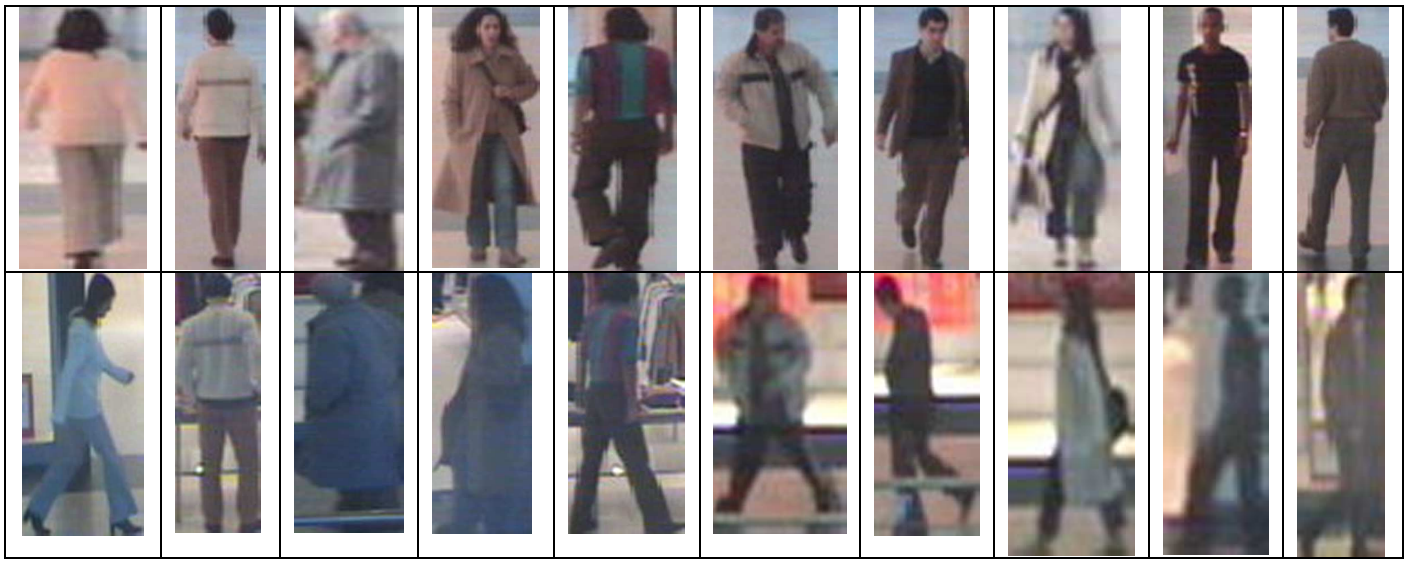

Figure 3: Typical low resolution views of the persons used as part of the models (top-line), and typical views of the same person in the other camera from which we try to re-identify.

The re-identification performance evaluation is done with the precision-recall metric:

$$\Pr ecision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}\, call = \frac{TP}{T \arg et\ number}$$

with TP (True Positives) = number of correct query-model re-identification matching, and FP (False Positives) = number of erroneous query-model matching.

The resulting performance, computed on a total of 760 query video sequences of 3 seconds, is presented in table 1, and illustrated on a precision-recall curve on figure 5. The main tuning parameter of our method is the "score threshold", which is the minimum number of matched points between query and model required to validate a re-identification. As expected, it can be seen that increasing the matching score threshold, increases the precision but at the same time lowers the recall. Taking into account the relatively low image resolution, our person re-identification performances are good, with for instance 82% precision and 78% recall when the score threshold is set to a minimum of 15 matching interest points between query and model. Note that we here consider only the best match (if matching score over threshold).

For comparison, the best result reported in [6] on a set of videos including 44 different persons is 60% correct best match (and they achieve 80% only when considering if true match is included in the 4 best matches). It is of course difficult to draw any strong conclusion from that, as the video sets are completely different, and ours contain only 10 different persons v.s. 44 in work presented in [6].

Score threshold for query-model matching (number of matched points)	**Precision (%)**	**Recall (%)**
40	**99**	**49**
35	**97**	**56**
30	**95**	**64**
25	**90**	**71**
20	**85**	**75**
15	**82**	**78**
10	**80**	**79**
5	**80**	**80**

Table 1: precision and recall, as a function of the score threshold for query-model matching (ie minimum number of similar interest points).

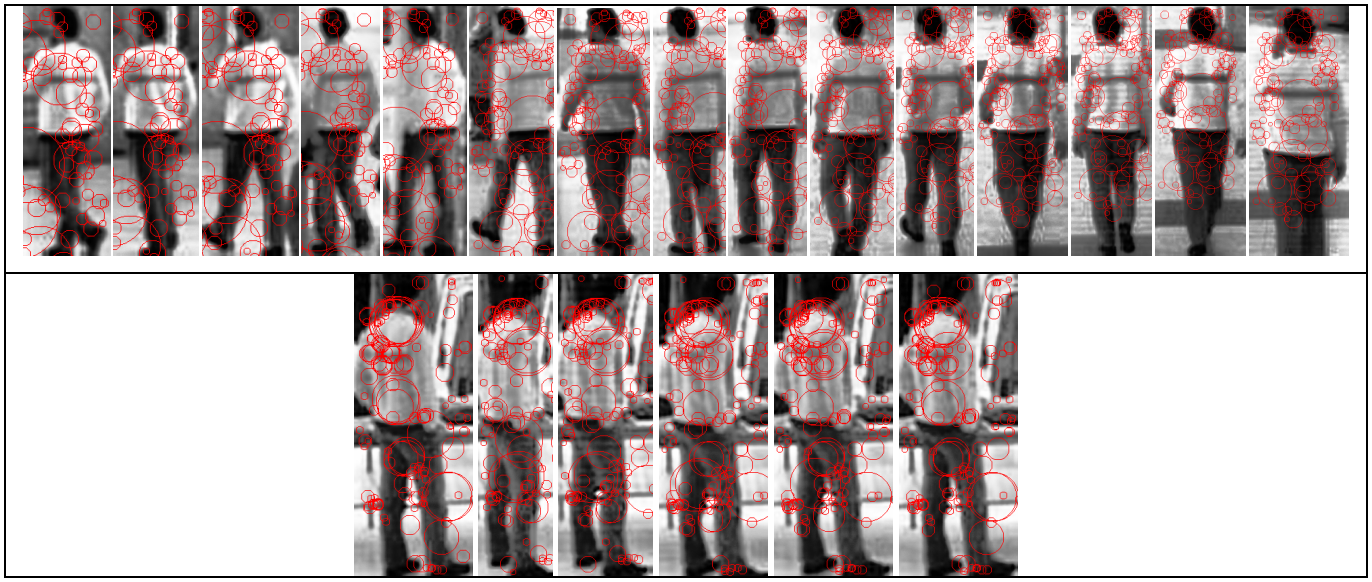

Figure 4:Visualization of detected key-points on 15 of the 21 images for one person's model (top line), and on the 6 images of a successfully matched re-identification query for the same person (bottom-line).

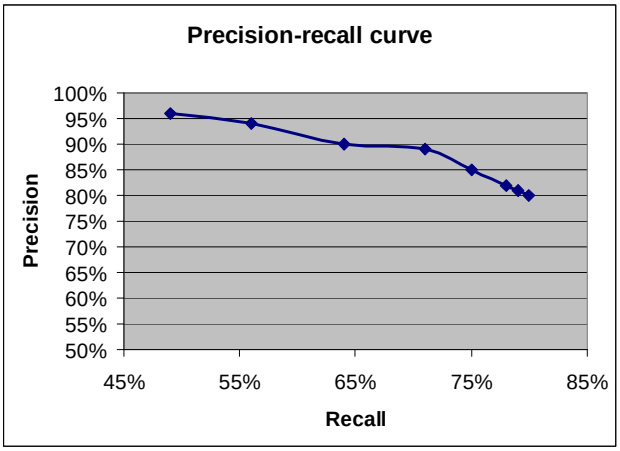

Figure 5:Precision-recall curve in our first person re-identification experiment.

The high execution speed of our re-identification method should also be emphasized: the computing time is less than 1/8 s per query, which is negligible compared to the 3 seconds necessary to collect the six images separated by ½ s.

More importantly, due to the logarithmic complexity of the KD-tree search with respect to the number of stored descriptors, the query processing time should remain very low even when large number of person models are stored. In order to verify this, we compared the re-identification computation time when varying the number of images used in model sequences, as reported in table 2. Indeed figure 6 shows that the re-identification computation time scales logarithmically with number of stored descriptors. Since the number of stored descriptors is roughly proportional to the number of images used, if 100 to 1000 person models were stored instead of 10 (with ~ 20 images for each), the KD-tree would contain 10 to 100 times more key-points, i.e. ~ 0.25 to 2.5 millions of descriptors. Extrapolating a little from figure 6, we therefore expect a query computation time < 1/5 s for a re-identification query among hundreds of registered person models. However, the *reliability* of re-identification among such a high number of persons remains to be verified.

Number of images used in model sequences	Total number of stored interest points	Computation time for re-identification (ms)
1	**1117**	**123**
2	**2315**	**132**
4	**5154**	**141**
8	**11100**	**149**
16	**22419**	**157**
24	**32854**	**161**

Table 2: total number of interest points, and re-identification computation time when varying the number of images used for the model for each stored person

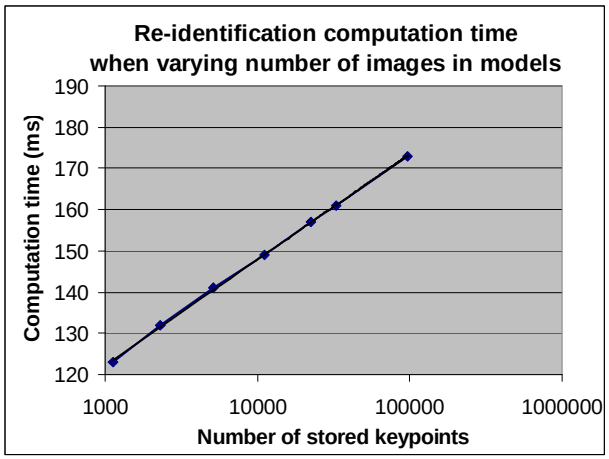

Figure 6:Re-identification computation time as a function of stored key-point descriptors; the dependence is clearly logarithmic

4. CONCLUSIONS AND PERSPECTIVES

We have presented a new person re-identification approach based on matching of interest-points collected in query short video sequence with those collected in longer model videos used for each previously seen person.

Our first experiments on low-resolution videos have shown very promising inter-camera person re-identification performances (a precision of 82% for a recall of 78%).
It should be noted that our matching method is very fast, with typical computation time of 1/8s for re-identification of one target person among 10 stored signatures for previously seen persons in other cameras. Moreover, this re-identification time scales logarithmically with the number of stored person models, so that the computation time would remain below 1/5 second for a real-world-sized system potentially involving tracking of hundreds of persons.

More thorough evaluations have to be done and are currently under progress, including on other video corpus, and especially with more registered persons, in order to assess re-identification reliability among large number of persons. Also, our re-identification scheme will soon be integrated in the global video-surveillance processing, which will allow to restrict interest points inside the person area, therefore excluding most background key-points, which should improve significantly the performance of our system.
Finally, we also hope to further increase performances, either by exploiting relative positions of matched interest points, and/or by applying a machine-learning to the set of "models" built for respective registered persons.

5. REFERENCES

[1] Tu P., Doretto G., Krahnstoever N., Perera A., Wheeler F., Liu X., Rittscher J., Sebastian T., Yu T., and Harding K., "An intelligent video framework for homeland protection" *Proceedings of SPIE Defence and Security Symposium - Unattended Ground, Sea, and Air Sensor Technologies and Applications IX*, Orlando, FL, USA, April 9--13, 2007.

[2] Park U., Jain A., Kitahara I., Kogure K., and Hagita N., "ViSE: Visual Search Engine Using Multiple Networked Cameras", *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)-Volume 03*, 1204-1207 (2006).

[3] Pham T., Worring M., and Smeulders A., "A multi-camera visual surveillance system for tracking reoccurrences of people", *Proc. of 1st AC/IEEE Int. Conf. on Smart Distributed Cameras*, Vienna, Austria, 25-28 sept. 2007.

[4] Lantagne M., Parizeau M., and Bergevin R. , "VIP : Vision tool for comparing Images of People", *Proceedings of the 16th IEEE Conf. on Vision Interface*, pp. 35-42, 2003

[5] Arth C., Leistner C., and Bishof H., "Object Reacquisition and Tracking in Large-Scale Smart Camera Networks**"**, *Proc. of 1st AC/IEEE Int. Conf. on Smart Distributed Cameras* held in Vienna, Austria, 25-28 sept. 2007.

[6] Gheissari N., Sebastian T., and Hartley R., "Person Reidentification Using Spatiotemporal Appearance", *Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'2006)*-Volume 2, IEEE Computer Society, pp. 1528-1535, New-York, USA, June 17-22, 2006.

[7] Sivic J. and A. Zisserman A., "Video Google: A text retrieval approach to object matching in videos", *Proceedings of 9th IEEE International Conference on Computer Vision (ICCV'2003)*, held in Nice, France, 11-17 october 2003.

[8] Lowe D., "Distinctive Image Features from Scale-Invariant Keypoints" *International Journal of Computer Vision, Vol. 60*, pp. 91-110, *Springer*, 2004,

[9] Bay H., Tuytelaars T., and Gool L. V., "SURF:Speeded Up Robust Features", *Proceedings of the 9th European Conference on Computer Vision (ECCV'2006)*, Springer LNCS volume 3951, part 1, pp 404--417, 2006

[10] Beis J. and Lowe D., "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces", In *Proc. 1997 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1000-1006, Puerto Rico, 1997.

[11] Gandhi T. and Trivedi M., "Person Tracking and Reidentification: Introducing Panoramic Appearance Map (PAM) for Feature Representation", *Machine Vision and Applications: Special Issue on Novel Concepts and Challenges for the Generation of Video Surveillance Systems*, August 2007.

[12] Yilmaz A., Javed O., and Shah M., "Object Tracking: A Survey", *ACM Journal of Computing Surveys*, Vol. 38, No. 4, 2006.

[13] Basharat A., Zhai Y., Shah M., "Content Based Video Matching Using Spatiotemporal Volumes", *Computer Vision and Image Understanding (CVIU)*, Volume 110, Issue 3, Pages 360-377, June 2008.