Interest points harvesting in video sequences for efficient person identification

Omar Hamdoun, Fabien Moutarde, Bogdan Stanciulescu, and Bruno Steux
Robotics Laboratory (CAOR), Mines ParisTech, 60 Bd St Michel, F-75006 Paris, FRANCE
{Omar.Hamdoun,Fabien.Moutarde,Bogdan.Stanciulescu,Bruno.Steux}@ensmp.fr

Abstract

We propose and evaluate a new approach for identification of persons, based on harvesting of interest point descriptors in video sequences. By accumulating interest points on several sufficiently time-spaced images during person silhouette or face tracking within each camera, the collected interest points capture appearance variability.
Our method can in particular be applied to global person re-identification in a network of cameras. We present a first experimental evaluation conducted on a publicly available set of videos in a commercial mall, with very promising inter-camera pedestrian re-identification performances (a precision of 82% for a recall of 78%). 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 ~ 1/5s second.
Finally, we also present a first feasibility test for on-the-fly face recognition, with encouraging results.

1. Introduction and related work

In many video-surveillance applications, it is desirable to determine if a presently visible person, vehicle, or object, has already been observed somewhere else in the network of cameras. This kind of problem is commonly known as “re-identification”, and a general presentation of this field for the particular case of person tracking 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, and also, for global persons, rapid change in clothes appearance.

A first category of person re-identification methods rely on biometric techniques (such as face or gait recognition). Face identification in “cooperative” context on high-resolution images with well-controlled pose and illumination can now be done with very good performance (see eg [12] or [13]). However, on-the-fly face matching in wide-field videos is still a very challenging problem.

A second group of methods try to perform person re-identification using no biometrics but only global appearance. Among these, various approaches have been proposed: signature based on color histograms (such as in [2], [3] or [14]), texture characteristics (see eg [4]). 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].

We here propose and evaluate a re-identification scheme using matching of interest points harvested in several images during short video sequences. The central point of our algorithm lies in the exploitation of image sequences, 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]. However, contrary to 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], but use a direct matching between interest point descriptors.

We hereafter describe our method in more details, present a first performance evaluation on publicly available real-world videos, and a preliminary feasibility test for application to on-the-fly face identification.

2. Description of our re-identification scheme

In this section we detail the algorithmic choices made in our re-identification approach. Our method can be separated in two main stages: a learning phase, and a recognition phase. The learning consists in taking advantage of tracking of a given person, vehicle or object in a sequence from one camera, in order to extract interest points and their descriptors necessary to build the model.

The interest point detection and descriptor computation is done using "key-points" functions available in the Camellia image processing library (http://camellia.sourceforge.net). These Camellia key-points detection and characterization functions are implementing a very quick variant, which shall be described elsewhere in more details, inspired from SURF [9] but even faster. SURF itself is an extremely efficient method (thanks to use of integral images) inspired from the more classic and widely used interest point detector and descriptor SIFT [8].

Apart from some technical details, the main difference between Camellia keypoints and SURF lies in optimization of Camellia implementation, using integer-only computations, which makes it even faster than SURF, and particularly well-suited for embedding in camera hardware.

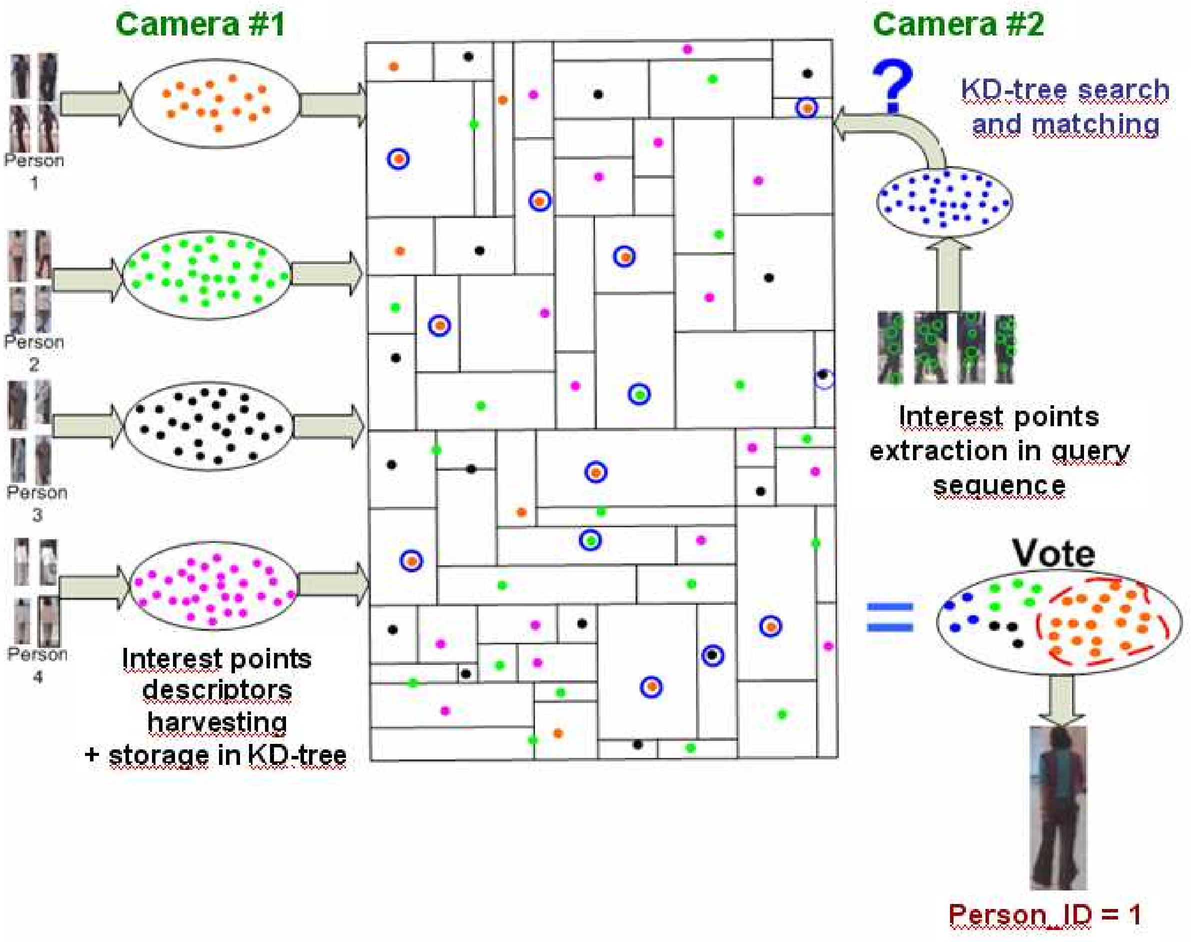
Camera #1
Person 1
Person 2
Person 3
Person 4
Interest points descriptors harvesting + storage in KD-tree
Camera #2
?
KD-tree search and matching
Interest points extraction in query sequence
Vote
=
Person_ID = 1

Figure 1: schematic view of model building (left), and of re-identification of a query (right).

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 model 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 model 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 for application to re-identification of persons

For the person re-identification application, a first experimental evaluation of our method has been conducted, 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.

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 2). Camera color potential variability is avoided by working in grayscale. Illumination invariance is ensured by histogram equalization of each person's bounding-box.

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

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

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

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

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

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

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 3. 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.

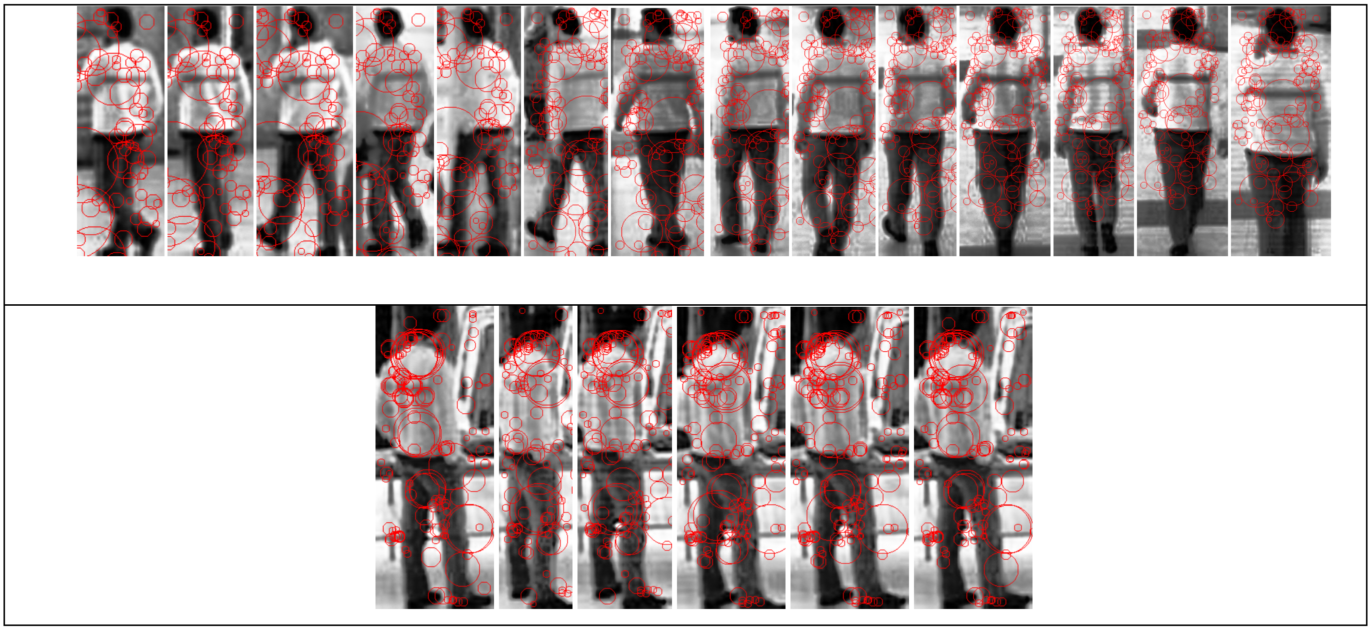

Figure 2: Visualization of detected key-points on 14 of the 21 images for one person's model (top line), and on the 6 images of a successfully matched re-identification query for the same person (bottom-line).

For comparison, [14] which use color histograms for re-identification on another video set including 15 different persons report best results of ~80% positive rate for ~80% true negative proportion among negatives. Also, 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 these comparisons, as the video sets are completely different, with different numbers of persons, which may be more or less similar, but it seems that the order of magnitude of our first evaluation of re-identification performances is rather encouraging.

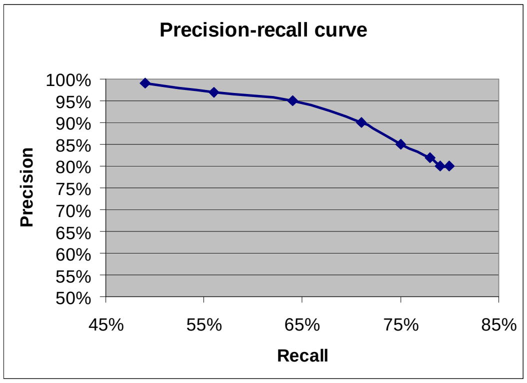

Figure 3: Precision-recall curve in our first person "global" re-identification experiment

In order to quantify the advantage of harvesting interest points on several images, we tried to use various numbers of images per model of person, and checked the impact on obtained precision and recall. As can be seen on figure 4, there is a very clear and significant *simultaneous* improvement of both precision *and* recall when the number of images per model is increased. This validates the interest of the concept of harvesting interest points for the same person on various images.

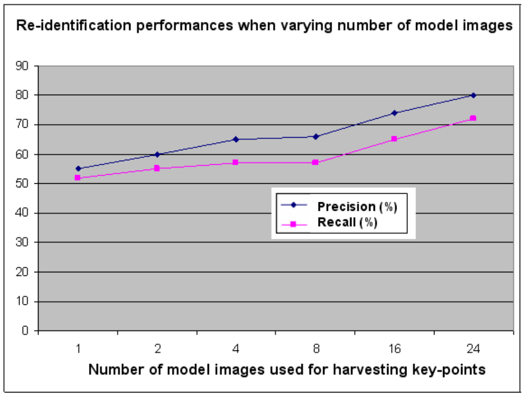

Figure 4: Influence of the number of images used per model on the re-identification precision and recall

It is also important to emphasize the high execution speed of our re-identification method: 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.

Table 2: Total number of interest points, and re-identification computation time as a function of the number of images used for each person model

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

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 if large number of person models were 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.

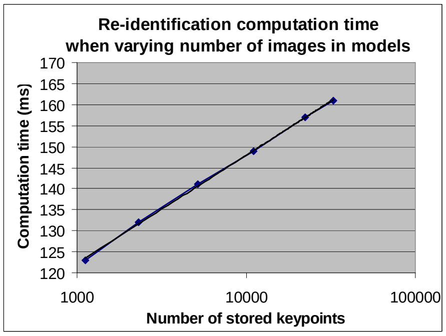

Figure 5: Re-identification computation time as a function of number of stored keypoint descriptors; the dependence is clearly logarithmic

Indeed figure 5 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 from figure 5, 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 of course remains to be verified.

4. Feasibility study of application to on-the-fly face identification

To evaluate the generality of our approach, we have also begun preliminary tests for a feasibility study of application to on-the-fly face identification face in real time. We have installed four IP cameras in our research lab in order to build up our own experiment with different cameras.

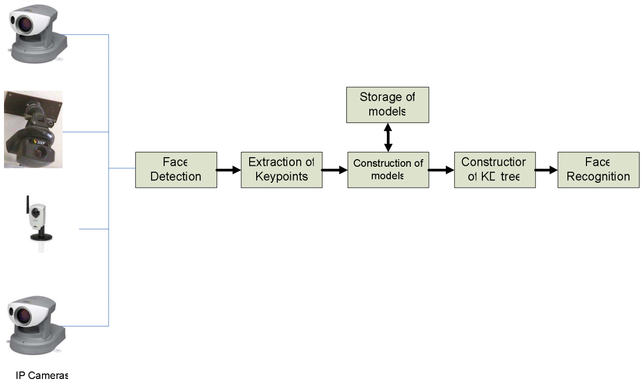

Figure 6: The general diagram of our application to real-time on-the-fly face identification

The general principle is the same as for person global re-identification, except that we need a face detection module to locate face for model harvesting as well as for re-identification by face. Figure 6 shows a general diagram of our on-the-fly face identification system.

For the face detection, we began by using just the standard Viola-Jones face detection algorithm [11] implemented in open source library OpenCV. Figure 7 shows the points of interest extracted by the Camellia keypoints detector on the faces of several persons in our lab. As for global person re-identification, we use several sufficiently different images for building the model of each person.

The quantification of the identification performances of our system is currently still in progress, but we have already qualitatively interesting results, as illustrated on figure 8 where one can see some successful on-the-fly face identification, including some on difficult situations with occlusion, or when a person wears dark sunglasses (while no such image example was present in its model).

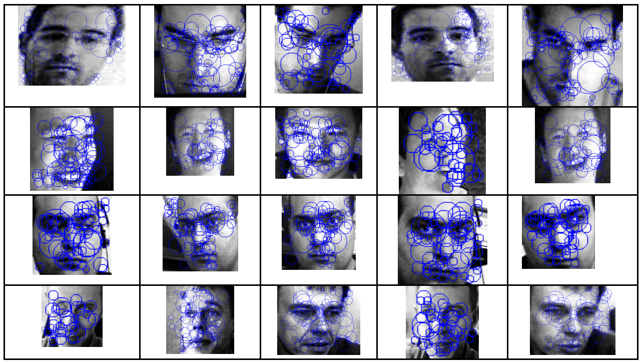

Figure 7: Examples of interest points extracted with Camellia KeyPoint detector inside detected faces used as models

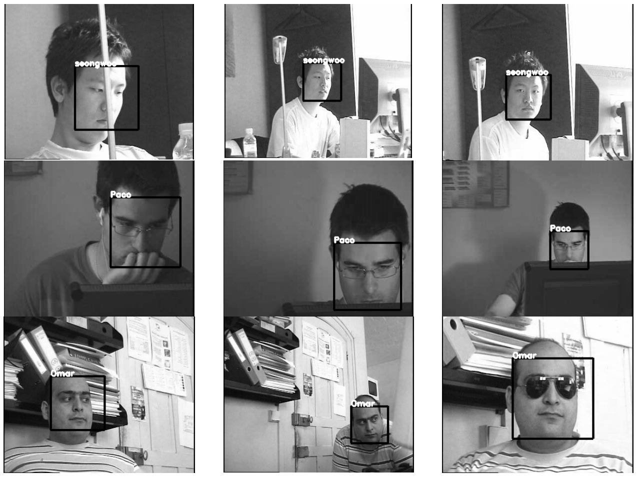

Figure 8: Examples of successful on-the-fly face identification with our interest point harvesting approach

5. Conclusions

We have presented a new re-identification approach based on matching of interest-points collected in query short video sequence with those harvested in longer model videos used for each previously seen and registered "object" (person, face or vehicle).

We have conducted a first evaluation for application to global pedestrian re-identification in multi-camera system on low-resolution videos. This yielded 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.

We have also set up a first feasibility evaluation of application of the same method for on-the-fly face identification, with encouraging successful face identification in difficult situations (occlusion, adding sunglasses, etc…).

More thorough evaluations have to be done for pedestrian re-identification, including on other video corpus, and with more registered persons, which are currently under progress. For on-the-fly face recognition, a quantitative evaluation is underway. 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 keypoints, 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 built "models".

6. References

[1] Tu, P.; Doretto, G.; Krahnstoever, N.; Perera, A.; Wheeler, F.; Liu, X.; Rittscher, J.; Sebastian, T.; Yu, T. & 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. & 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.; Smeulders A., "A multi-camera visual surveillance system for tracking reoccurrences of people", *Proc. of 1^{st} AC/IEEE Int. Conf. on Smart Distributed Cameras* held in Vienna, Austria, 25-28 sept. 2007.

[4] Lantagne, M.; Parizeau, M. & 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.; Bishof H., "Object Reacquisition and Tracking in Large-Scale Smart Camera Networks**"**, *Proc. of 1^{st} AC/IEEE Int. Conf. on Smart*

Distributed Cameras held in Vienna, Austria, 25-28 sept. 2007.

[6] Gheissari, N.; Sebastian, T. & 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] Herbert Bay, Tinnr Tuytelaars & 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. & 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] Viola, P. & Jones, M., 'Rapid object detection using a boosted cascade of simple features', *Proc. CVPR* **1**, 511—518, 2001.

[12] Belhumeur P. N., Hespanha J., and Kriegman D.J., “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection”. In *IEEE European Conference on Computer Vision*, pages 45–58, 1996

[13] Draper B., Baek K., Bartlett M.S., and Beveridge R., “Recognizing faces with PCA and ICA”. *Computer Vision and Image Understanding: Special issue on Face Recognition*, 91, 2003

[14] J. Orwell, P. Remagnino, G.A. Jones, "Optimal Color Quantization for Real-time Object Recognition" in *'Real Time Imaging'*, 7(5) Academic Press, October, pp. 401-414. ISBN/ISSN 1077-2014 (2001).