Probabilistic Matching Pair Selection for SURF-based Person Re-identification

Mohamed Ibn Khedher, Mounim A. El-Yacoubi, Bernadette Dorizzi

Department of Electronics and Physics
Institut Mines-Télécom: Télécom SudParis
Evry, France
{mohamed.ibn_khedher, mounim.el_yacoubi, bernadette.dorizzi}@it-sudparis.eu

Abstract: The objective of this paper is to study the performance of human reidentification based on multi-shot SURF and to assess its degradation according to the angular difference between the test and reference video scene view angles. In this context, we propose a new automatic statistical method of acceptance and rejection of SURF correspondence based on the likelihood ratio of two GMMs learned on the reference set and modeling the distribution of distances resulting from matching sequences associated with the same person and with different persons respectively. The experimental results show that our approach compares favorably with the state of the art and achieves a good performance.

1 Introduction

Human re-identification has been a fast evolving research topic over the last years because of its diverse applications (subway stations, hospitals, shopping centers, etc.). Re-identification is an important video surveillance task. In a camera network, and given two cameras possibly having different scene views, if a person leaves the view of one camera and reappears in the other, the re-identification system must be able to re-identify him/her and continue monitoring. The performance of a human re-identification system can be affected by several factors such as the variability of illuminations and diversity of viewing angles: the subject may look different due to change in camera viewing angles and lighting conditions.

Since re-identification is a recent research topic, a few existing studies are found. Methods of re-identification can be single-shot based (one image is used to build a person signature [GT08, ZGX09]) or multiple-shot based (multiple images are exploited to build a person signature [GSH06, Ham10]).

Overall, methods of re-identification can be classified into two main approaches: appearance approaches and local approaches. Appearance approaches commonly used in the state-of-the-art include color and texture. These features may be combined for obtaining a more representative descriptor [MTCar].

Color is the most used appearance primitive, usually in the form of histograms, cumulative histograms which are invariant to scale [BSM+10] or dominant color description. Among color features, the Dominant Color Descriptor (DCD) [BCBT10] and the Major Color Spectrum Histogram Representation (MCSHR) [MCP07] compute the recurrent RGB color values used to represent a patch. [FBP+10] considers Maximally Stable Color Regions (MSCR) representing a person by patches having a homogenous color. Among texture features, [BCBT10] used the AdaBoost scheme to find out the most discriminative Haar-like feature textural information set for each individual. [BSM+10] used the ratios of colors, ratios of oriented gradients and ratios of saliency as textural features.

The second category of re-identification approaches is based on local features. It consists of representing an image by local interest points. Several interest points detectors have been considered (Harris [HS88], Harris-Laplace [MS04], and Fast Hessian [BETVG08]). For interest point description, Scale Invariant Features Transformation (SIFT) [Low01], Speed-Up Robust Features (SURF) [BETVG08], Shape Context [BMP02] and Gradient Location and Orientation Histogram (GLOH) [MS05] are used. Interest points are employed in different fields such as object recognition [Low01], face recognition [BLGT06] and pedestrian detection [SLMS05]. Geissari et al. [GSH06] extracted spatio-temporal interest points described by color and structural information. Arth et al. [ALB07] used the PCA-SIFT for re-identification in large networks of cameras. Hamdoun et al. [Ham10] performed person re-identification by matching SURF interest points extracted at each frame and accumulated through short video sequences and a KD-tree was used in order to speed up the matching process. [JA10] used SIFT to build an Implicit Shape Model (ISM). In [JA11], person re-identification is performed in three stages. In each of the first two, matching serves as a filtering stage for the following one. A Bag of Words of SIFT is used in the first stage. Spatial information is added in the second stage and SIFT are used directly in the third stage. In this last stage, only filtered SIFT are used for matching. Moreover, the authors detected the angle view of test sequence and applied a mirror transformation to SIFT descriptors in order to convert the test view feature description into one closer to training sequence feature description.

One of the main observations regarding state of the art re-identification methods is that there is no approach systematically outperforming the others: each approach has its own strengths and limitations. Color features are easy to extract and to exploit. The problem of these features is that people may wear similar clothes and in this case color features will be insufficient to discriminate people. However, they are useful when combined with other features (Haar-Wavelets and DCD in [BCBT10]). In addition, these features are sensitive to camera parameters and illumination conditions. In fact, differences in illumination cause measurements of object colors to be biased towards the color of the light source [Kvi11]. Texture-based approaches also may extract easily a rich information (encode information of the entire frame). One of the main drawbacks of these approaches is their sensitivity to the camera view. On the other hand, Interest Points are detected in a way they are invariant to scale (like SIFT) or scale and rotation (like SURF). For two images temporally close, we expect that the detected points have close positions in both images. Visually, however, it is not the case, since interest points are not stable.

The objective of this paper is to study the performance of person re-identification from video sequences and its degradation depending on the reference and test view angles as well as on their differences. To this end, we consider a re-identification scheme based on local interest points, namely SURF, because of their relative robustness towards camera view angle change. To overcome the instability of these points, we follow the multi-shot re-identification approach using all images in order to increase the reproducibility of interest points between two similar video scenes. Several works based on the multi-shot approach have been considered [Ham10, BCP$^+$10, BCBT11, BCPM12]). This work is different from [Ham10, BCP$^+$10, BCBT11, BCPM12] where each person is represented by few frames.

To design a re-identification system based on interest points, the matching step is crucial. Hamdoun et al. [Ham10] choose an empirically preset number of best matched points between query and reference and use them in a majority vote scheme to validate a re-identification. In [dOdSP09], two interest points p_0 and p_1 ($p_0 \in$ Reference and $p_1 \in$ Query) are matched if $d(p_1, p_0) < c\ d(p_1,p_i)\ \forall\ p_i \in$ Reference, where c is a preset coefficient $c<1$ and $d(.,.)$ is the Euclidian distance. The two approaches above estimate a threshold, empirically, during the matching process. In this work, we propose a new method that avoids such an empirical setting and automatically accept/reject an interest point pair according to a statistical modeling of distances between interest points.

The method is based on two Gaussian Mixture Models (GMMs) modeling the distribution of distances between interest points associated with same person in the first case, and with different persons in the second case. The decision to accept or reject a pair is then taken upon whether or not the likelihood ratio between the two GMMs above is higher or lower than one. To train the two GMMs, the reference video sequences can be conveniently exploited as it will be explained in Section 2.3.

This paper is structured as follows. In section 2, we present the major stages of our re-identification system. The experimental results of our approach are given in section 3. A conclusion and a perspective are finally presented.

2 Person Re-identification System Description

Our approach basically consists of four stages: 1) Detection of the Region Of Interest (ROI), 2) Feature extraction using SURF, 3) Interest Points matching and filtering using GMMs and 4) human re-identification based on majority vote rule. The following figure (Figure 1) shows the flowchart of our approach.

2.1 Detection of the ROI

Our approach takes as input the region of interest containing the human silhouette. The detection of the ROI can be based on background subtraction or on a machine learning method such as the one proposed by Dalal et al. taking as input HOG descriptor [DT05].

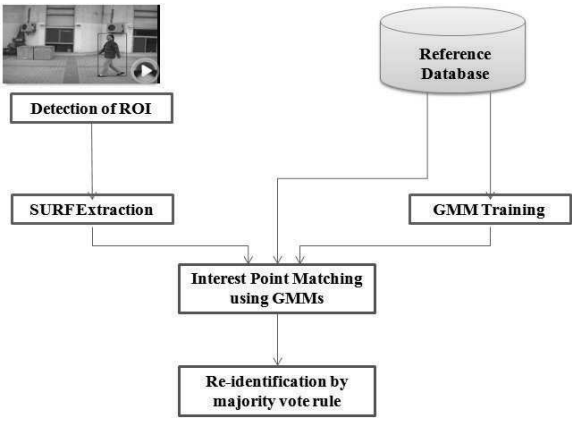

Figure 1: Re-identification stages

In the database used for our experiments (CASIA-A Database), the background subtraction result is available. Figure 2(a) shows a screenshot of an original image; Figure 2(b) shows its binary silhouette and Figure 2(c) shows its ROI.

Figure 2: Figure 2: a) original image, b) binary silhouette, c) ROI image

2.2 Feature Extraction

State of the art shows that many interest points' detectors and descriptors are used. Each method differs from the others in terms of description, invariance criteria and running time. Some evaluations of different Interest Points show conflicting results. In [Ham10], SURF outperforms SIFT. However, in [BS11], GLOH and SIFT outperforms both Shape Context and SURF. Our own evaluation of SIFT and SURF on CASIA-B database [CAS01] showed that SURF outperforms SIFT. Hence, SURF is used in this paper. SURF descriptors [BTG06] are the accompanying descriptors of the fast-hessian interest points' detectors [BTG06]. A SURF descriptor is computed as a sum of local intensity differences within a 4x4 grid around the interest point. These intensity differences are calculated as responses to first-order Haar-Wavelets. For illumination invariance, the descriptor is normalized to unit length. Figure 3 shows the detected SURF points within a frame ROI.

Figure 3: Feature extraction

2.3 Correspondence Acceptation/Rejection based on GMM

To develop a robust interest point pair matching for re-identification, it is important to set a mechanism that is able to automatically discard any matched pair likely to be associated with two *different* persons and to accept any matched pair likely to be associated with the *same* person. To this end, we consider two Gaussian Mixture Models (GMMs): GMM_1 modeling the distribution of distances between interest points associated with the same person in the first case, and GMM_2 modeling the distribution of distances between interest points associated with different persons in the second case. The decision to accept or reject a pair with distance *d* is then taken upon whether or not the likelihood ratio *LR* between the two GMMs above is higher or lower than one.

$$LR = \frac{P(d/GMM_1)}{P(d/GMM_2)}$$

The pair is retained if *LR* >1 and is discarded if *LR* ≤ 1. As shown, this decision mechanism is automatic and no empirical threshold setting is needed.

GMMs is a probabilistic model which can approximate any distribution, given a sufficient number of mixture components. It is a special case of the generative graphical Model HMM (Hidden Markov Model) [Rab89] having one state. For the case where the covariance matrices are diagonal, the associated probability density function is defined as follows:

$$p(x) = \sum_{g=1}^{G} c_g \prod_{f=1}^{F} \frac{1}{\sqrt{2\pi\sigma_{gf}^2}} \exp\left(\frac{-1}{2}\left(\frac{x_f - \mu_{gf}}{\sigma_{gf}}\right)^2\right)$$

where *G* is the number of mixture components, c_g is the weight associated with component

g, F is dimension of the feature vector x, and μ_{gf} and σ_{gf} are the mean and standard deviation of x with respect to component g. GMMs are trained using the Expectation-Maximization (EM) algorithm [DLR77].

In our case, both GMM_1 and GMM_2 are univariate and model distribution distances between matched interest points. The likelihood of a distance d with respect to each GMM is:

$$p(d/GMM_i) = \sum_{g=1}^{G} c_{gi} \frac{1}{\sqrt{2\pi\sigma_{gi}^2}} \exp\left(\frac{-1}{2}\left(\frac{d-\mu_{gi}}{\sigma_{gi}}\right)^2\right) \qquad i = 1;2$$

where μ_{gi} and σ_{gi} are the mean and standard deviation of component g for GMM_i.

To train the two GMMs, the reference video sequences can be conveniently exploited as explained below. The training of the two GMMs is performed on a reference database consisting of 20 persons.
Each person provides one sequence for each of the six viewing angles (Figure 4). For each combination of viewing angles in the reference database ($angle_1$, $angle_2$) (36 combinations) and for each person "P_1" from the reference database, we apply the following two steps:

1. Matching 1: We consider the two sequences of the person *"P_1"*, corresponding to the two viewing angles ($angle_1$ and $angle_2$). Then, interest points of these two sequences are matched and distances of correspondences are added to set S_{same} consisting of the distances resulting from matching sequences from the same person.

2. Matching 2: We choose, randomly, another person *"P_2"* from the reference database and we select the sequence of *"P_2"* with view angle *"$angle_2$"*. Distances resulting from the interest points' matching between the sequence of person *"P_1"* with view angle *"$angle_1$"* and the sequence of person *"P_2"* with view *"$angle_2$"* contribute to set S_{diff} consisting of distances resulting from matching sequences of two different persons.

This strategy raises one issue: when $angle_1$ and $angle_2$ are equal, only one reference sequence is available and Matching 1 is no longer possible. To overcome this issue, we divide, in this case, the sequence into two sub-sequences, one containing the odd frames and the other containing the even frames (in order to keep the temporal aspect of the video) and perform matching between them in order to generate the data (distances) contributing to S_{same}. Although the number of interest points to be matched is halved in average, this scheme allows us to estimate conveniently the distribution of distances resulting from matching sequences of the same person with the same view angle even though only one reference sequence is available.

After sets S_{same} and S_{diff} are generated in this way, the parameters of GMM_1 and GMM_2 are estimated by considering a number of mixture components that is close to the number of the view angles available in the re-identification dataset.

2.4 Human Re-identification

Human re-identification basically consists of three stages. 1) matched pairs' selection, 2) GMMs decision and 3) majority vote decision rule. Given a test sequence and a reference database, the objective is to assess whether a sequence from the same person as the test sequence is within the reference database. The first step, matched pairs selection, consists on the following scheme: for every interest point from the test sequence, the closest point in the reference database is determined. In the second step, matched interest points are filtered using the GMM-based likelihood ratio criterion as detailed in section 2.3. In the third stage, the retained interest point pairs are submitted to the majority vote decision rule. For each retained pair, a vote is added to the person associated with the reference interest point. The person obtaining the majority of votes is claimed as the re-identified person.

3 Experiments and Results

3.1 Database

CASIA Dataset-A [CAS01] was created in 2001 and consists of 20 persons. Each person has 12 image sequences, 4 sequences for each of the three directions, i.e. parallel, diagonal and frontal to the image plane. This database is suitable for evaluation of multi view re-identification because it includes people moving in the scene along 6 different directions: 0°, 90°, 135°, 180°, 270°and 315°(see Figure 4).

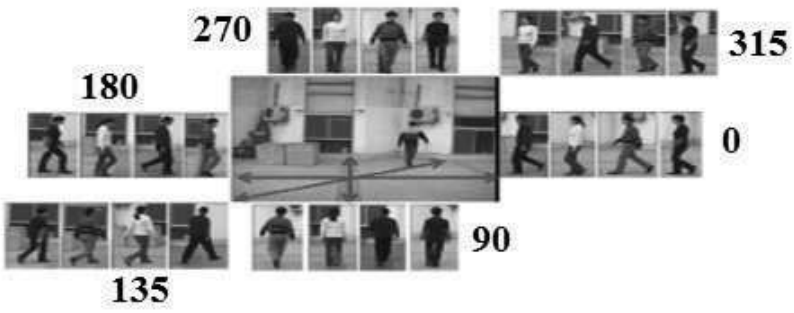

Figure 4: The CASIA-A dataset

3.2 Test Protocol and Configuration

For evaluation, two sequences are available for each walking direction. One is used as reference and the other as test. For GMM-based likelihood ratio decision, two GMMs are

learned on the reference database, as detailed in section 2.3; the number of Gaussians is 5 for each GMM. Re-identification performance is presented with the Correct Classification Rate (CCR) defined by the ratio of the number of persons correctly recognized over the total number of tested persons.

3.3 Results

In this section, we present the results of our approach based on SURF matching and correspondence selection based on GMM likelihood ratio. Evaluation is performed for every possible combination (test view, reference view). Since the database contains 6 test view angles and 6 reference view angles, 36 experiments are performed. Results are shown in Figure 5 and Table 1. Figure 5 shows re-identification performance according to the angular difference between test view and reference view, regardless of the actual test and reference view angles. Table 1 shows results for different combinations of reference views (columns) and test views (rows).

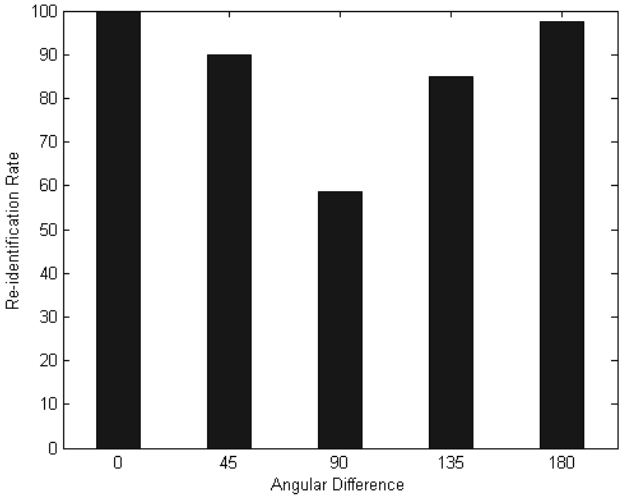

Figure 5: Re-identification performance according to the angular difference between test view and reference view on the CASIA-A Database

Figure 5 shows that re-identification is nearly perfect when test and reference views are identical or symmetric (angular difference = 180°). This proves the robustness of SURF matching for re-identification when reference and test sequences are similar. Moreover, re-identification performance decreases with increasing angular difference. Performance is better when test and reference share some visible part. For instance, considering the angle "0°" as a reference angle, Table 1 shows that the best results are obtained for test view "0°" and "180°" (100% and 100%) then results decrease slightly with test view "135°" and "315°" (90% and 100%) and then more significantly with test view "90°" and "270°"

Table 1: Correct Classification Rate for different combinations of view angles on the CASIA-A database.

Angle	**0°**	**90°**	**135°**	**180°**	**270°**	**315°**
0°	100	40	90	100	35	90
90°	75	100	100	75	90	95
135°	90	70	100	100	70	100
180°	100	50	95	100	30	85
270°	70	95	95	75	100	95
315°	100	60	100	85	70	100

(75% and 70%). This result is compatible with the example in Figure 4 where we can see that view "0°" is similar to view "180°" and that it shares a common part with views "135°" and "315°"; the shared part then decreases even further with view "90°" and "270°".

Figure 6 compares our results with those of the approach in [JA11] based on Implicit Shape Model and SIFT features (to the best of our knowledge, this is the only work reporting results on the same dataset for human re-identification).

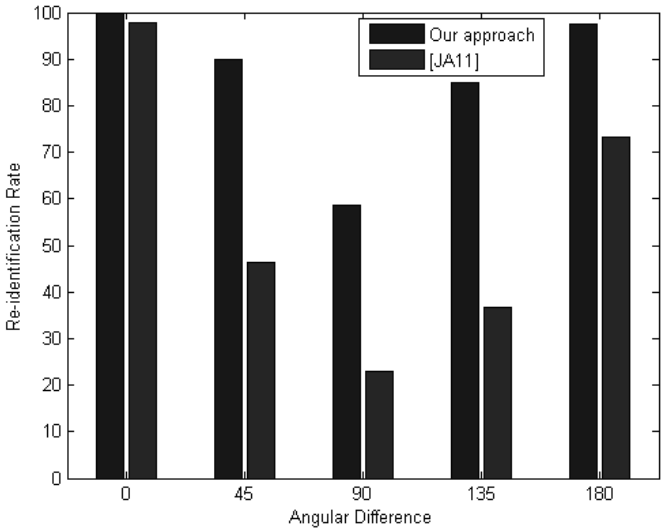

Figure 6: Re-identification performance comparison

Figure 6 shows that our method significantly outperforms [JA11], especially when reference and test view angles are different. This shows the robustness of SURF matching under different view angles when the redundancy of whole reference and test video sequences is fully exploited for matching and a proper mechanism for rejecting/accepting matched interest point pairs is considered.

4 Conclusions and Perspectives

This paper has studied the performance of a multi-shot SURF based person re-identification system and assessed performance degradation according to the angular difference between test and reference video scenes. It proposed also an automatic method of acceptance and rejection of SURF correspondence based on likelihood ratio of two GMMs learned on the reference set, which avoids selection of matching SURF pairs by empirical means. The results obtained in our experiments show the relative robustness of SURF for different camera views when whole video sequences are exploited for effective interest point matching. Our approach compares favorably with the only one work found in the literature using the same database [JA11].
In the future, we will investigate the task of feature combination. Specifically, the aim will be to seek cooperation strategies of re-identification methods based on color, geometry of the shape and interest points matching in order to optimize the re-identification performance and/or select automatically the method suitable for each re-identification scenario.

References

[ALB07] Clemens Arth, Christian Leistner, and Horst Bischof. OBJECT REACQUISITION AND TRACKING IN LARGE-SCALE SMART CAMERA NETWORKS, 2007.

[BCBT10] Slawomir Bak, Etienne Corvee, Francois Bremond, and Monique Thonnat. Person Re-identification Using Haar-based and DCD-based Signature. *Advanced Video and Signal Based Surveillance, IEEE Conference on*, 0:1–8, 2010.

[BCBT11] Slawomir Bak, Etienne Corvee, François Bremond, and Monique Thonnat. Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid. In *Advanced Video and Signal-Based Surveillance*, Klagenfurt, Autriche, August 2011.

[BCP+10] Loris Bazzani, Marco Cristani, Alessandro Perina, Michela Farenzena, and Vittorio Murino. Multiple-Shot Person Re-identification by HPE Signature. In *ICPR*, pages 1413–1416, 2010.

[BCPM12] Loris Bazzani, Marco Cristani, Alessandro Perina, and Vittorio Murino. Multiple-shot person re-identification by chromatic and epitomic analyses. *Pattern Recognition Letters*, 33(7):898–903, 2012.

[BETVG08] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[BLGT06] Manuele Bicego, Andrea Lagorio, Enrico Grosso, and Massimo Tistarelli. On the use of sift features for face authentication. In *In: Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW). (2006*, page 35. IEEE Computer Society, 2006.

[BMP02] S. Belongie, J. Malik, and J. Puzicha. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, April 2002.

[BS11] M. Bauml and R. Stiefelhagen. Evaluation of local features for person re-identification in image sequences. *Advanced Video and Signal Based Surveillance, IEEE Conference on*, 0:291–296, 2011.

[BSM+10] Guy Berdugo, Omri Soceanu, Yair Moshe, Dmitry Rudoy, and Itsik Dvir. Object Reidentification in Real World Scenarios across Multiple Non-overlapping Cameras. In *Proc. of the 18th European Signal Processing Conference (EUSIPCO 2010*, Aalborg, Denmark, Aug 2010.

[BTG06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *In ECCV*, pages 404–417, 2006.

[CAS01] CASIA. http://www.cbsr.ia.ac.cn/english/Gait

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[dOdSP09] Icaro Oliveira de Oliveira and Jose Luiz de Souza Pio. People Reidentification in a Camera Network. *Dependable, Autonomic and Secure Computing, IEEE International Symposium on*, 0:461–466, 2009.

[DT05] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR (1)*, pages 886–893, 2005.

[FBP+10] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010.

[GSH06] Niloofar Gheissari, Thomas B. Sebastian, and Richard Hartley. Person Reidentification Using Spatiotemporal Appearance. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 1528–1535, Washington, DC, USA, 2006. IEEE Computer Society.

[GT08] Douglas Gray and Hai Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 262–275, Berlin, Heidelberg, 2008. Springer-Verlag.

[Ham10] Omar Hamdoun. *Détection et ré-identification de piétons par points d'intérêt entre caméras disjointes*. PhD thesis, École Nationale Supérieure des Mines de Paris, 2010.

[HS88] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.

[JA10] Kai Jungling and Michael Arens. Local Feature Based Person Reidentification in Infrared Image Sequences. *Advanced Video and Signal Based Surveillance, IEEE Conference on*, 0:448–455, 2010.

[JA11] Kai Jüngling and Michael Arens. View-invariant Person Re-identification with an Implicit Shape Model. In *2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillancei (AVSS)*, page 6, Aug. 2011.

[Kvi11] Igor Kviatkovsky. Color Invariants For Person Re-Identification. Master's thesis, Technion - Israel Institute of Technology, 2011.

[Low01] David G. Lowe. Local Feature View Clustering for 3D Object Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 682–688. Springer, 2001.

[MCP07] Christopher Madden, Eric Dahai Cheng, and Massimo Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Mach. Vision Appl.*, 18:233–247, May 2007.

[MS04] Krystian Mikolajczyk and Cordelia Schmid. Scale & Affine Invariant Interest Point Detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[MS05] Krystian Mikolajczyk and Cordelia Schmid. A Performance Evaluation of Local Descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27:1615–1630, October 2005.

[MTCar] R. Mazzon, S.F. Tahir, and A. Cavallaro. Person re-identification in crowd. *Pattern Recognition Letters*, To appear.

[Rab89] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[SLMS05] Edgar Seemann, Bastian Leibe, Krystian Mikolajczyk, and Bernt Schiele. An evaluation of local shape-based features for pedestrian detection. In *In Proc. BMVC*, 2005.

[ZGX09] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating Groups of People. In *BMVC*, 2009.