# Evaluation of Local Features for Person Re-Identification in Image Sequences

Martin Bäuml[1]
[1]Institute for Anthropomatics
Karlsruhe Institute of Technology
Adenauerring 2, 76137 Karlsruhe, Germany
baeuml@kit.edu

Rainer Stiefelhagen[1,2]
[2]Fraunhofer Institute of Optronics,
System Technologies and Image Exploitation
Fraunhoferstr. 1, 76131 Karlsruhe, Germany
rainer.stiefelhagen@kit.edu

## Abstract

*In this paper we present a comparative study of local features for the task of person (re-)identification. A combination of state-of-the-art interest point detectors and descriptors is evaluated. The experiments are performed on a novel dataset which we make publicly available for future research in this area. The results indicate that there are significant differences between the evaluated descriptors, with GLOH and SIFT outperforming both Shape Context and SURF descriptors. The evaluated interest point descriptors perform equally well, with a slight advantage for the Hessian-Laplace detector. The Harris-Affine and Hessian-Affine affine invariant region detectors do not provide any performance advantage and therefore do not justify their additional computational expense.*

## 1. Introduction

Person *re-identification* has attracted a lot of research attention in recent years. For many applications it is not necessary to actually uniquely identify a person, it suffices to determine previous or future occurrences of the same person in other images or image sequences. As such it can serve as building block in person tracking for connecting tracks over blind gaps between multiple cameras or occlusions, in person retrieval to search for specific persons of interest in multimedia data or surveillance footage, or for short-term identification of persons surveillance camera network.

Since unique identification is not required for person re-identification, it is prudent to take other than biometric features into account, which often are unreliable in uncontrolled environments. Many recent approaches utilize the whole-body appearance of a person based on the assumption that it does not change significantly within a relevant time-frame and thus is well suited for re-identification. In fact, full body appearance is also very well exploited by humans [7]. For a recent overview over appearance-based person re-identification approaches the reader is referred to [6].

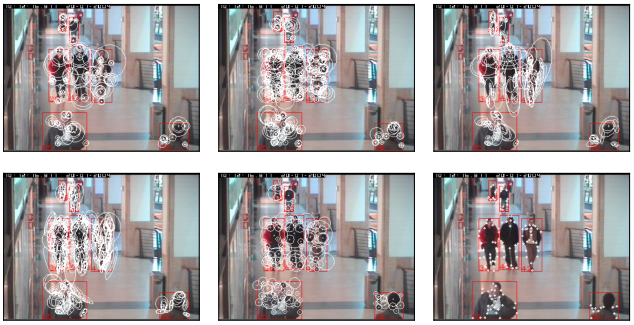

Figure 1: Responses from different interest point detectors. From top left to bottom right: Harris, Harris-Laplace, Hessian-Laplace, Harris-Affine, Hessian-Affine and Fast-Hessian.

We will focus here on person re-identification approaches based on *local features* [8, 10, 13] (in this paper, we will use the term local features as synonym for local interest points in combination with local descriptors). Local features have shown to be able to successfully establish correspondences between related images. Accordingly, they have been utilized among others in image retrieval [20], object recognition [15], pedestrian detection [22], person tracking [12] and face recognition[4]. We will briefly review some local feature-based approaches to person re-identification in the following.

With a focus on real-time performance, Hamdoun et al. [10] extract SURF features [2] from video frames in intervals of 0.5 seconds. Features are matched efficiently using kd-trees. A simple voting model is employed for closed-set recognition. Gheissari et al. [8] use the Hessian affine invariant interest point operator [16] to locate interest points. The local region around an interest point is described by an HSV-edgel descriptor. Two interest points between two images are matched if one is the nearest neighbor of all interest points in the other's image and vice versa. A final validation step further prunes false correspondences.

Jüngling et al. [13] build upon a SIFT-based person tracking approach [12, 15] for person re-identification in infrared images. Instead of finding nearest neighbors of the features directly, features are matched to visual words, which are learned beforehand. Two person tracks are compared by individually comparing features that match to the same visual word.

While all of the above approaches are basically independent of the actual local feature type, in their implementation and evaluation they all focus on a single feature type. However, previous evaluations of local features suggest that not every local feature type is equally suitable for a given task (e.g. [17, 18, 22]). The main aim of this paper is to determine which features are most suited for person re-identification. Our contributions are the following: (i) We perform a comparative study of state-of-the-art local features for open-set person re-identification. (ii) We propose a simple approach to person re-identification using local features, exploiting multiple connected frames from a person tracker if available. (iii) We present a novel dataset for person re-identification with properties unavailable in previous datasets to further encourage research in this area.

## 2. Local Features for Person Re-Identification

For the evaluation, we use local features for person re-identification in the following way: First, we detect interest points in frames where a person is present (we assume that we have a person tracker which provides us with a rough bounding box around the person). We then compute features for all interest points that lie within a persons bounding box. A person model is trained from one or multiple sample sequences. We model a person as a *bag of features*, i.e. we collect the set of extracted features without any additional information about their spatial layout within the image. For the identification of a new person, we match the extracted features to all previously trained person models and compute scores based on the distances of the features to the models. If multiple frames from a person track are available, we fuse the scores from the individual frames to achieve a better identification.

We will now first introduce the evaluated interest point detectors and descriptors, and then explain the training and testing of the person models in more detail.

### 2.1. Local Interest Point Detectors

For the nomination of interest points, we evaluate six state-of-the-art interest point detectors. Some example detections are visualized in Figure 1.

**Harris** The Harris corner detector [11] detects image structures with a high *cornerness* such as corners and T-junctions. Harris and Stephens define a cornerness measure which is large if both eigenvalues of the second moment matrix are simultaneously large, i.e. when there are strong intensity changes in orthogonal directions at a given point. Interest points are selected at local maxima of the cornerness function. Harris corners are invariant with respect to translation and rotation but not to scale changes.

**Harris-Laplace** The Harris-Laplace detector [17] adds scale invariance to the Harris detector. For this, a scale-adapted second moment matrix is used, i.e. the local derivatives are calculated at different coarse scale levels. Local maxima of the Harris cornerness function (now based on the scale-adapted second-moment matrix) nominate interest point candidates. A *characteristic scale* is determined for each interest point candidate by finding a local extremum over scale of the Laplacian-of-Gaussian response at that point. Candidates without a significant local extremum in scale-space are discarded.

**Hessian-Laplace** interest points [17] are very similar to Harris-Laplace interest points. However, the detection of interest point candidates is based on the determinant of the scale-adapted Hessian matrix, where a local maximum corresponds to a blob-like structure, i.e. a round or ellipse-shaped intensity pattern.

**Harris-Affine, Hessian-Affine** The affine invariant versions of both Harris and Hessian detectors [16] aim at achieving invariance with respect to arbitrary affine transformations. After finding interest points at characteristic scales, the shape of a characteristic elliptical affine region around the interest point is determined in an iterative way. This is done by repeatedly estimating the shape of the affine region based on the second moment matrix, then transforming the region to a circle, until conversion.

**Fast-Hessian** interest points [2] are based on an approximate version of the Hessian matrix, efficiently calculated from integral images without the need for a scale-space image pyramid. The determinant of the approximate Hessian is used for both interest point and characteristic scale selection by searching for local 3D maxima. The detected image structures are similar to the ones detected by the Hessian-Laplace detector. The detector is not affine invariant.

### 2.2. Local Descriptors

With the success of local features in computer vision, a great number of local descriptors have been proposed. We focus here on some of the most prominent ones.

**SIFT** The scale invariant feature transform (SIFT) descriptor [15] is computed as a histogram of the gradient distribution in the region around a detected interest point. The gradient's orientation is quantized to 8 orientation bins, its location to one of $4 \times 4$ square regions, resulting in a 128-dimensional descriptor. The descriptor is normalized in order to obtain illumination invariance.

**Shape Context (SC)** is an edge-based descriptor. Edges

are computed using the Canny edge detector [5]. The descriptor consists of a histogram over the edge points, taking into account the location in 9 log-polar bins and edge orientation in 4 bins. The resulting descriptor is an extended version of the original Shape Context descriptor [3] and has 36 dimensions.

**Gradient Location and Orientation Histogram (GLOH)** descriptors [18] combine ideas from both SIFT and shape context. The descriptor is computed from gradients as in SIFT, but the location binning is performed in a log-polar manner similar to shape context. With 17 location bins and 16 orientation bins the intermediate descriptor has 272 dimensions, which are reduced to 128 dimensions using PCA.

**Speeded-up Robust Features (SURF)** descriptors [2] are the accompanying descriptors to the fast-hessian interest point detector. It is computed as sums of local intensity differences within a $4 \times 4$ grid around the interest point. These intensity differences are calculated as responses of first-order Haar-Wavelets which can be computed very efficiently on arbitrary scales using integral images. For illumination invariance the descriptor is normalized to unit length.

### 2.3. Bag-of-Features Person Model

We model a person's appearance using a *bag-of-features* representation, i.e. we describe it as a collection of local parts, ignoring their spatial (and for videos also their temporal) structure. This simple model has first been used for text retrieval, but also successfully been adapted to object recognition (e.g. [23]) and person re-identification [8, 10]. We chose it for its simplicity and the ability to evaluate the local features performance without any influence of a spatial model such as in [13]. Of course it can be expected that adding spatial information improves the overall results, but this shall not be our focus here.

Given a set of training images for a number of persons, we build one bag of features for *each* person by extracting all local features covering the person in the training images. The person's location in the image is determined from labeled ground truth data. We use ground truth instead of the output of a person tracker in order to be independent of tracking failures in our evaluation.

The obtained person models allow us to find a test feature's nearest neighbor with respect to each of the trained persons separately. For a test image, we compute the distance of all features within the person's bounding box to each of the person models by summing up the distances of all test features to their respective nearest neighbors in the person models:

$$d_i(k) = d_i(\mathcal{X}_\mathbf{k}) = \sum_j^{|\mathcal{X}_\mathbf{k}|} d\left(\mathbf{x}_j, \mathrm{NN}_i\left(\mathbf{x}_j\right)\right) \quad , \quad (1)$$

where $\mathcal{X}_\mathbf{k}$ is the set of local features in test frame $k$, $\mathbf{x}_j$ is the $j$-th feature in $\mathcal{X}_\mathbf{k}$, and $\mathrm{NN}_i\left(\mathbf{x}_j\right)$ is the nearest neighbor of $\mathbf{x}_j$ to any local feature in the model of person $i$. The assumption behind this scoring method is that a local feature from an unseen test image is more similar to a feature from the same person (i.e. the distance to the nearest neighbor is smallest) than to a feature from a different person.

Obviously, we need to find a lot of nearest neighbors in large sets of local features. In order to make this computationally tractable, we approximate the nearest neighbor search by using kd-trees which in our experiments speeds up the search by one to two orders of magnitudes compared to the naïve brute-force *linear scan* method. We will show that the speedup comes with basically no penalty in recognition performance (cf. Figure 3).

### 2.4. Normalization and Temporal Fusion

In camera networks we usually acquire videos instead of still images. A person tracker can therefore provide multiple, temporally connected instances of a person as a track.

In order to determine the identity of a person using a whole track of test frames, we first compute the model distances for each of the track's frames individually as described in Section 2.3 and then perform sum-rule fusion [14] over all track frames (Eq. 4). Since every frame's person bounding box can contain a different number of features, it is not beneficial to combine the frame-based distances directly but to normalize them first (Eq. 2 and 3). In detail, the person scores for a track are calculated from the individual frame distances as follows:

*1. Min-max-normalization* of the model distances to the interval $[0, 1]$. For each frame, the lowest model distance for the frame $min(d_i(k))$ is mapped to 1, the highest distance $max(d_i(i))$ to 0, and all remaining distances linearly between 1 and 0 according to

$$s_i(k) = \frac{d_i(k) - min(d_i(k))}{max(d_i(k)) - min(d_i(k))} \quad , \quad (2)$$

where $s_i(k)$ is the resulting raw frame score for person $i$ in frame $k$. Besides making distances between different frames comparable, this also has the nice property of turning distances into scores in a parameter-less way.

*2. L1-Normalization* of the obtained scores, i.e. so that their sum equals 1:

$$s_i^*(k) = \frac{s_i(k)}{\sum_i s_i(k)} \quad , \quad (3)$$

*3.* The *fusion* is performed by averaging the normalized scores over the whole track (sum-rule fusion):

$$s_i^{seq} = \frac{1}{N} \sum_k^N s_i^*(k) \quad . \quad (4)$$

The normalization by the length of the track $N$ is necessary for open-set recognition. Since the decision whether

Figure 2: Example frames of 30 of the 61 labeled persons from our person re-identification dataset.

the person is known or unknown is based on whether the best sequence score $s_i^{seq}$ is higher or lower than a threshold $\theta$, shorter tracks would otherwise be biased towards the unknown class.

## 3. Performance Evaluation

For the evaluation we use a subset of the publicly available CAVIAR dataset[1]. The dataset shows people walking down a corridor in a Lisbon shopping center. The resolution of the 26 clips is $384 \times 288$ pixels with a frame rate of 25 frames per second. We labeled the identities of 61 different persons and extracted 281 tracks using the provided bounding box labels from the original dataset[2].

Among the 61 persons are actually some who changed clothes between different clips. We labeled those as two different persons, since our goal is person identification from full-body appearance under the assumption that people do not change their clothes significantly between training and recognition. In order to obtain a larger number of tracks per person, we divided in some cases one longer track into multiple tracks of the same person with at least a 10 frame gap between the tracks. See Figure 2 for examples of the extracted persons. This dataset overcomes some shortcomings of the few other publicly available datasets for person re-identification since it contains videos instead of still images (opposed to [9, 24]) and actually multiple, different tracks of a large number of persons (as opposed to [21] where there is only one track of each person).

For the computation of interest points and descriptors, we use the implementations of Mikolajczyk[3] and Bay et al.[4]. For the approximate nearest neighbor search we use the FLANN library [19]. The number of kd-trees in all experiments conducted in this paper was set to 32 and training precision to 0.95.

### 3.1. Baseline

In order to show that a local feature-based approach justifies the additional computational expense, we also compare it to the performance obtained by describing a person's appearance by color histograms, which is by far the most widely used method due to its simplicity, and robustness against articulation changes, for example as appearance model for person tracking [1].

For this baseline method, we compute RGB color histograms from the bounding box region labeled in the data. Each channel is divided in 8 bins, resulting in a $8 \times 8 \times 8 = 512$ dimensional descriptor. From the color histograms we similarly build bag-of-feature person models as described in Section 2.3, i.e. each person model consists of the histograms extracted from all frames in the training tracks.

### 3.2. Evaluation Criteria

We perform the evaluation on the task of *open-set* person re-identification. An open-set classifier first needs to decide whether a person has been seen in the training set or is *unknown*. If a person is classified as known, we further determine the identity among the trained persons. We can evaluate the recognition performance in terms of False Acceptance Rate (FAR), Correct Classification Rate (CCR) and False Classification Rate (FCR), defined as

$$FAR = \frac{\text{\#false acceptances}}{\text{\#unknown samples}} = \frac{\left|\left\{\mathcal{C}\left(x_k^{-1}\right) = S_i \ : \ i > 0\right\}\right|}{|X_{\text{unknown}}|}$$

$$CCR = \frac{\text{\#correct classific.}}{\text{\#known samples}} = \frac{\left|\left\{\mathcal{C}\left(x_k^{i}\right) = S_i \ : \ i > 0\right\}\right|}{|X_{\text{known}}|}$$

$$FCR = \frac{\text{\#false classific.}}{\text{\#known samples}} = \frac{\left|\left\{\mathcal{C}\left(x_k^{i}\right) \neq S_i \ : \ i > 0\right\}\right|}{|X_{\text{known}}|},$$

where we denote the sets of known and unknown test sequences as

$$X_{\text{known}} = \left\{x_k^i | i \in 1, \ldots, n\right\},$$
$$X_{\text{unknown}} = \left\{x_k^i | i = -1\right\}.$$

and our open-set classifier as a function

$$\mathcal{C}\left(\boldsymbol{x}\right) = S_i, \quad i \in \{-1, 1, \ldots, n\}.$$

### 3.3. Temporal Fusion, Normalization and NN Approximation

In this section we will briefly discuss the influence of the usage of videos over single frames, the normalization and the effect of the nearest neighbour approximation. The results presented in this section are based on Hessian-Laplace interest points ($t = 200$) in combination with the GLOH descriptor.

The min-max-normalization in combination with the subsequent L1-normalization provides a significant increase

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1.

[2] We will make our identity and track labels available for download at http://cvhci.anthropomatik.kit.edu/projects/pri.

[3] http://www.robots.ox.ac.uk/~vgg/research/affine/

[4] http://www.vision.ee.ethz.ch/~surf/

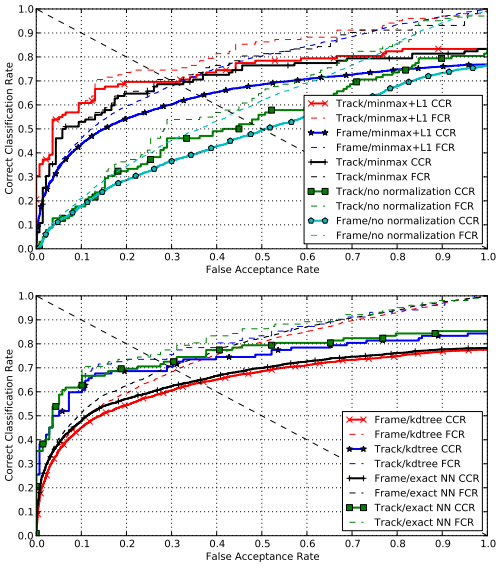


Figure 3: (top) Normalization and track-based recognition (*Frame/** denote single frame results, while *Track/** denote results after fusion.) (bottom) Recognition performance comparison of exact Nearest Neighbor computation and approximate Nearest Neighbor computation.

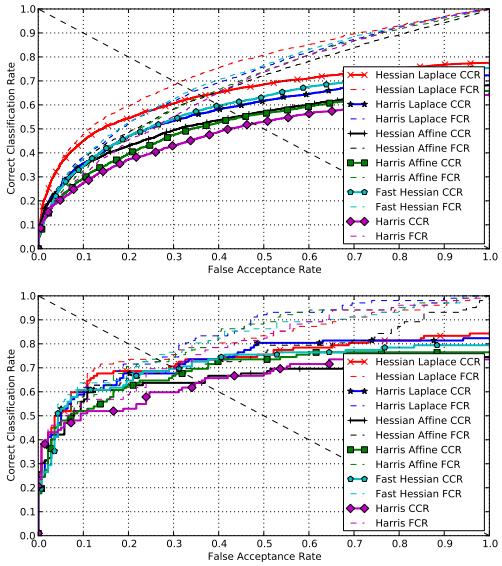


Figure 4: Comparison of frame-based (top) and track-based detector performance (bottom) for different interest point types.

in recognition performance for both frame- and track-based identification (cf. Figure 3(top)). Track-based identification with normalization outperforms the frame-based classification significantly due to the additional robustness gained by the fusion over time.

Using kd-trees instead of brute-force linear scan for nearest neighbour search, we achieved a speed-up of one to two orders of magnitude, resulting in an average classification time *per track* of 1.75 seconds compared to 65.5 seconds for the linear scan. The approximation does not have any significant impact to the recognition performance (cf. Figure 3(bottom)).

### 3.4. Evaluation of Interest Point Detectors

We will now investigate the performance of the different interest point detectors from Section 2.1. As descriptor we use the GLOH descriptor (which we will show in the next section is quite suitable for that task). We consistently used a detection threshold of $t = 200$, yielding a good coverage for all interest point types (cf. Figure 1).

Figure 4 shows the frame-based and track-based results (both with normalization). While the frame-based results indicate quite a clear advantage of the Hessian-Laplace interest point detector, after track-level fusion there is no clear outperformer. On track-level they perform equally well between around $60\%$ and $70\%$ correct classification rate at equal error rate (EER). The slight underperformance of Harris can be explained by the lack of scale invariance of the Harris detector, since training and test images of a person can be of largely different sizes when people are walking along the corridor. This result is consistent with other experiments which also showed a disadvantage of the Harris detector in the presence of scale changes [17].

More surprising is the fact that the affine version of the detectors cannot achieve a clear performance advantage. One could have expected that an affine invariant detector could better handle articulation variations of a walking person. One reason could be that the variations are too irregular to be found consistently by an affine invariant detector. The low resolution of the images could also render the benefits of an affine approximation of the transformation of a region around an interest point useless. Their additional computational effort therefore cannot be justified.

### 3.5. Evaluation of Interest Point Descriptors

Since there was no clear advantage of any of the interest point types, we performed the experiments for the descriptor evaluation with the Hessian-Laplace detector. Figure 5 shows the results for both frame- and track-based recognition with normalization. The gradient-based descriptors GLOH and SIFT significantly outperform the other two descriptors and both achieve a recognition performance of around $70\%$ CCR at EER. Their histogram binning seems to be able to best cope with the non-rigid deformation of the human body. The shape context descriptor also displays a remarkable performance, given its low dimensionality compared to SIFT and GLOH. The SURF descriptor achieves only around $52\%$ CCR at EER.

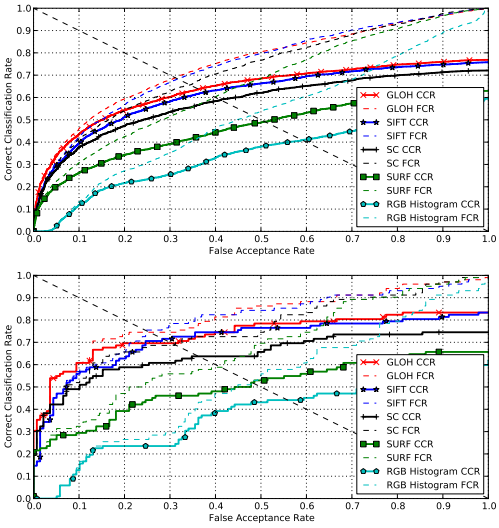


Figure 5: Comparison of descriptor performance. Frame-based (top) and track-based classification (bottom).

## 4. Conclusion

In this paper we have presented an evaluation of local features for person re-identification. We found that none of the tested state-of-the-art interest point detectors provides a significant performance advantage. The Harris corner detector performed slightly below average, due to its missing scale invariance. Surprisingly, affine region detectors did not outperform the scale invariant detectors, therefore their additional computational requirements cannot be justified. Within the set of tested interest point descriptors, GLOH and SIFT outperformed SC and SURF, achieving around 70% CCR at EER.

The performance differences between different types of descriptors underline the need for comparative studies as we conducted in this paper. Despite recent advances, person re-identification using local features remains challenging, which might in part be due to the fact that the current descriptors describe mainly shape and texture. We will explore in future research if extending local features to color can overcome some of the problems.

## 5. Acknowledgments

We thank Krystian Mikolajczyk and Herbert Bay for making their interest point detector and feature descriptor implementations available. We acknowledge the CAVIAR dataset, created in the EC funded project/IST 2001 37540. This project has been partially funded by the German Federal Ministry of Education and Research (BMBF) under project PaGeVi.

## References

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *CVIU*, 2008.

[3] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4), 2002.

[4] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli. On the use of SIFT features for face authentication. *CVPR Workshop on Biometrics*, 2006.

[5] J. Canny. A computational approach to edge detection. *PAMI*, 8(6), 1986.

[6] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Ambient Intelligence and Humanized Computing*, 2011.

[7] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008.

[8] N. Gheissari, T. B. Sebastian, and R. Hartley. Person Reidentification Using Spatiotemporal Appearance. *CVPR*, 2006.

[9] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS Workshop*, 2007.

[10] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. *ICDSC*, 2008.

[11] C. Harris and M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, 1988.

[12] K. Jüngling and M. Arens. Detection and tracking of objects with direct integration of perception and expectation. *ICCV Workshop on Visual Surveillance*, 2009.

[13] K. Jüngling and M. Arens. Local Feature Based Person Reidentification in Infrared Image Sequences. *AVSS*, 2010.

[14] J. Kittler. Combining classifiers: A theoretical framework. *Pattern Analysis & Applications*, 1(1):18–27, Mar. 1998.

[15] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 2004.

[16] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. *ECCV*, 2002.

[17] K. Mikolajczyk and C. Schmid. Scale & Affine Invariant Interest Point Detectors. *IJCV*, 2004.

[18] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10), 2005.

[19] M. Muja. Fast approximate nearest neighbors with automatic algorithm configuration. In *Conference on Computer Vision Theory and Applications*, 2009.

[20] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *PAMI*, 19(5), 1997.

[21] W. R. Schwartz and L. S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. *Brazilian Symposium on Computer Graphics and Image Processing*, 2009.

[22] E. Seemann, B. Leibe, K. Mikolajczyk, and B. Schiele. An evaluation of local shape-based features for pedestrian detection. In *BMVC*, 2005.

[23] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. *ICCV*, 2003.

[24] W. Zheng, S. Gong, and T. Xiang. Associating groups of people. In *BMVC*, 2009.