A Soft-Biometrics Dataset for Person Tracking and Re-Identification

Arne Schumann, Eduardo Monari
Fraunhofer Institute for Optronics, System Technologies and Image Exploitation
{arne.schumann, eduardo.monari}@iosb.fraunhofer.de

Abstract

In this work we present a new dataset for the tasks person detection, tracking, re-identification, and soft-biometric attribute detection in surveillance data. The dataset was recorded over three days and consists of more than 30 individuals moving through a network of seven cameras. Person tracks are labeled with consistent IDs as well as soft-biometric attributes, such as a description of the clothing, gender, or height. Persons in the video data alter their appearance by changing clothes or wearing accessories. A second, clothing specific ID of each track allows for the evaluation of re-identification with or without the presence of clothing changes. In addition to video and camera calibration data, we provide evaluation protocols, tools and baseline results for each of the four tasks.

1. Introduction

Person detection, tracking and recognition or re-identification are important tasks in computer vision which, particularly in low resolution surveillance scenarios, present a difficult challenge.

In most surveillance systems automation is an important factor in improving search times, reducing costs by allowing for a smaller number of operators, and even increasing performance by alerting operators to incidences they might otherwise have missed or noticed too late. The basis for most surveillance tasks is the detection of persons. Person tracking then allows to follow and analyze a person's trajectory within a camera's field of view. In order to facilitate a successful tracking of persons across multiple, possibly non-overlapping cameras, person re-identification is applied. Additional information such as soft-biometric attributes can help communicate details to human personnel as well as increase performance of the re-identification task.

The dataset presented in this work was recorded specifically to allow for evaluation of these four tasks. It contains many of the realistic challenges that approaches have to address when dealing with real world surveillance data. This includes noisy images, challenging lighting conditions, a multitude of viewing angles, non-overlapping cameras, small resolutions, varying framerates, and camera types.

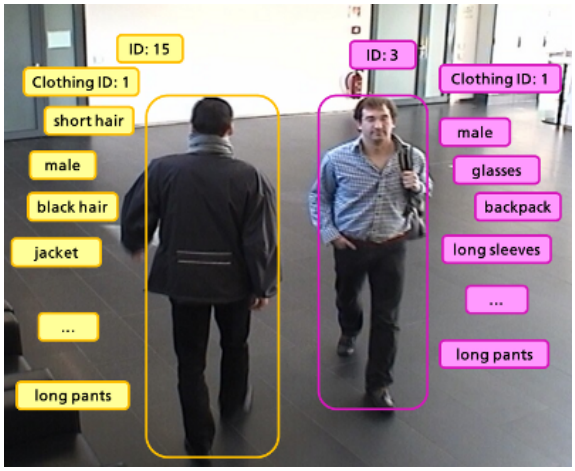

Figure 1. Impressions of the camera setup and annotations. Tracks labeled as bounding boxes with person IDs. Each track is annotated with a set of soft-biometric attributes and a more specific clothing ID which allows to differentiate between different clothing configurations of the same person.

Persontracks in the dataset are annotated with consistent IDs over all sequences which allows for the evaluation of person re-identification approaches. Additionally, annotations of soft-biometric attributes are provided for each track. This includes a description of an individual's clothes as well as attributes such as gender, ethnicity, height and hair color. Each track is annotated using two IDs, one global person ID and another which is specific to that person's clothing configuration. This makes it possible to evaluate re-identification approaches with or without the presence of clothing changes. An impression of the types of annotations in the dataset is depicted in Figure 1.

Our main contributions are as follows: 1) We provide a new, fully labeled video surveillance dataset for a unique combination of tasks, 2) subjects in the dataset change clothes and we provide a two-fold ID-labeling of tracks which allows for the evaluation of person re-identification with and without clothing changes, and 3) we provide many of the tools and baseline results presented in this work to make evaluation on the dataset as comfortable as possible.

The remainder of this work is structured as follows: Section 2 gives an overview of related datasets, Section 3 describes our new dataset and gives some statistics on video data and annotations. Section 4 presents baseline results for all of the main tasks the dataset was intended for and some concluding remarks are given in Section 5.

2. Related Work

A number of datasets already exist for person tracking and re-identification. We limit our discussion of existing datasets to those most established and most closely related to ours.

One of the most widely used datasets for person re-identification is the VIPeR dataset [11]. It offers images of 632 pedestrians walking through a network of two cameras. Its main limitation is that for each person only one image per camera is provided. The VIPeR dataset has recently been annotated with soft-biometric attributes [14]. The PRID dataset [13] contains short image sequences of 245 pedestrians recorded by two cameras. This dataset has been annotated with soft-biometric attributes as well [14].

Other re-identification data has been created from tracking datasets. The CAVIAR4REID dataset [7] contains sets of images for 72 pedestrians which were sampled from annotated person tracks in the CAVIAR dataset [1]. Similarly, the iLIDS-MA and iLIDS-AA datasets [2] were sampled from the iLIDS surveillance dataset [15]. They provide images for 40 individuals and 100 individuals, respectively. The iLIDS-AA images were created automatically by sampling the output of a person tracker. Another iLIDS-based re-identification dataset with an average of four images for 119 pedestrians was provided in [21]. The dataset presented in [19] provides crops from ground truth in the ETHZ tracking dataset [10] which was recorded by moving a camera on a mobile platform through an uncontrolled pedestrian area. The dataset contains multiple images of 146 pedestrians which were cropped from the same ground truth track.

All of these datasets have in common that they do not provide full or long tracks for individuals, offer few different camera views and are not suited for the evaluation of all steps in a surveillance system from person detection to multi-camera person tracking and re-identification.

More recently some datasets have been recorded that provide video and annotations for a full camera network and more closely represent a real surveillance scenario. The 3DPeS dataset [3] provides video from eight surveillance cameras placed in an outdoor courtyard. The videos contain 200 people shown on average in two different cameras. The CMV100 dataset [20] includes video data for 100 persons recorded by five cameras placed along an office hallway. In addition to person tracks and ids, this dataset also provides person foreground masks. Finally, in [5] an indoor surveillance dataset is provided which was recorded using eight cameras and contains 150 people.

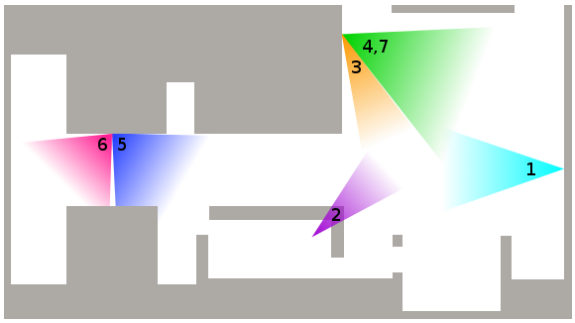

Figure 2. Floor plan of the area in which the dataset was recorded. The camera network consists of six IP cameras and one HD camera.

Our dataset falls in the latter category. It is recorded in a full camera network and suited to evaluate person detection, tracking and re-identification. Additionally, it provides labels to consider clothing changes for re-identification and labels to evaluate the accuracy of soft-biometric attribute detectors.

3. The Dataset

The dataset presented in this work was recorded using a camera network in the entrance area of our building [1]. The network consists of seven cameras. A floor plan with the positions and orientation can be seen in Figure 2 and the corresponding camera images are depicted in Figure 3. Six of the cameras are low resolution IP cameras and the seventh a camcorder. The camcorder was placed next to an IP camera in order to provide the same view at a higher resolution and framerate. It can be used to investigate the performance of an approach at different resolutions. The video data contains a number of challenges which are common in surveillance scenarios. Cameras can have different color footprints (e.g., cameras 4 and 7), some images are more noisy than others (e.g., camera 6) or images are strongly influenced by illumination (camera 5). Some of the cameras run at different frame rates and streaming the images through a network leads to occasional frame drops. However, we provide a tool to syncronously access images for all cameras. A summary of the most important camera parameters is given in Table 1. We provide calibration information for all seven cameras.

The recording took place over the course of three days. In total, 31 volunteers are visible in the dataset. Some participated on all three days, some only on single days. We recorded sequences of different degrees of difficulty for tracking and identification. In easier sequences (SS), persons walk alone through the camera network ensuring that they pass each camera at least twice. These sequences do

[1] sobisdata.org

Figure 3. Views from all seven cameras. Camera 4 and camera 7 have the same position and angle but different resolutions.

	Resolution	FPS	Type
Cam1	768×576	20	AXIS 214
Cam2	640×480	20	AXIS 211
Cam3	702×576	10	AXIS 223
Cam4	768×576	10	AXIS 214
Cam5	704×576	15	AXIS 211
Cam6	704×576	15	AXIS 211
Cam7	1440×1080	25	Camcorder

Table 1. Main characteristics of the cameras in the network.

not contain any occlusions and are well-suited to establish a reference performance or to be used as training data for appearance models, distance metrics, brightness transfer functions, travel times between cameras, or automatically learning the camera layout. Sequences of a moderate difficulty (SM) contain 3-4 persons walking through the camera network. Finally, challenging sequences (SG) contain up to 15 persons moving through the camera network in an uninstructed manner. These sequences contain a high degree of occlusions and ambiguity when persons cross from one camera to the next. Some impressions of the variety of sequences can be seen in Figure 6. In total, around five hours of video data was recorded for each camera.

The video data was manually annotated at every 5th frame and interpolated in between. Annotations include a bounding box for each person, a person ID which remains consistent over all three days and a second sub-ID which increments whenever a person changes their clothes. Examples can be seen in Figure 4. Clothing changes in the data range from small alterations such as wearing a backpack to complete clothing changes between different days. No instructions regarding how often and how drastically to

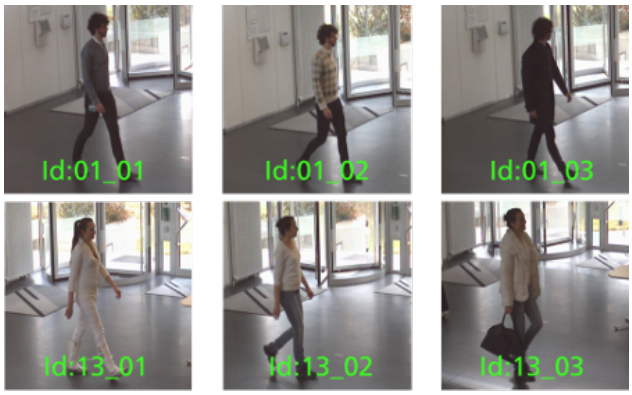

Figure 4. Two-fold person ids. The first value indicates the person, the second corresponds to the clothing configuration of that person.

change clothing configuration was given to the participants. On average, each person changed their appearance twice. In addition to bounding boxes and IDs, a number of soft-biometric attributes is annotated for each person. These include height, gender, age, ethnicity, hair color, hair style, beard, glasses, short sleeves, short pants, jeans, jacket, type of headwear, accessories, such as backpacks or suitcases, main upper body color, and main pants color.

4. Evaluation

We evaluated methods for each of the tasks the dataset is intended for to provide baseline results for future comparisons. Please refer to the dataset website for the most recent full sets of results for each of the tasks.

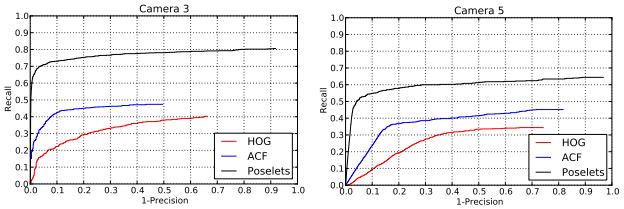

Figure 5. Precision-recall curves for the three person detectors. Detectors perform best in camera 3 while cameras with noise or illumination changes (such as camera 5) are more challenging.

4.1. Person Detection

Methods We used three well known approaches to detect persons. The holistic HOG person detector [8] is a popular choice for baseline results. We used a HOG detector with default settings as recommended in [8]. The detector was trained on the INRIA pedestrian dataset. The second holistic detector is based on Aggregate Channel Features (ACF) [9] and a boosting classifier. This detector was trained on the INRIA dataset as well. As a third option we used an implementation of the poselet person detector as described in [6]. This part-based detector uses the output of a large number of HOG-type body part detectors to merge into final person detections. We used the model provided by the authors which was trained on their own dataset (H3D [6]).

Each detector was run at every annotated (every 5th) frame of the video data. To keep the results as generally applicable as possible, no camera calibration information or groundplane model was used.

Results Person detections are evaluated by matching them against ground truth using the PASCAL VOC criterion for bounding box overlap thresholded at 0.5. Detector perfomance remains relatively constant over all video sequences but varies more drastically between cameras. Cameras with noise, illumination effects, or view angles the detectors were not trained for perform the worst (i.e., cameras 5 and 6). The best results are achieved using cameras 3 and 7. Figure 5 shows precision-recall curves for the best and worst performing cameras.

We provide the outputs and precision recall curves for all three detectors and all cameras as well as tools for evaluation on the dataset website.

4.2. Person Tracking

Methods The person tracker we applied follows the tracking-by-detection approach and is similar to the one described in [18]. The tracker consists of a particle filter combined with a simple histogram-based appearance model to avoid track-switches. We tracked persons within each camera separately. Tracking was performed in 2D image space and again no additional information such as groundplanes was used. As input we used the poselet person detections, due to their high accuracy.

	SS1	SS2	SM1	SG1
MOTA Camera 1	66.2	67.1	58.2	54.3
MOTA Camera 3	65.3	68.0	65.3	61.4
MOTA Camera 5	60.3	61.0	56.4	52.2

Table 2. Tracking accuracies for different cameras in easy (SS) and more challenging (SM,SG) sequences of the dataset.

Results We use the Multiple Object Tracking Accuracy (MOTA) [4] metric for evaluation of person tracks. MOTA combines the number of track misses $MISS_t$, the number of false positive tracks FP_t, the number of track switches MM_t and the number of ground truth tracks GT_t at time t into a single value:

$$MOTA = 1 - \frac{\sum_t MISS_t + FP_t + MM_t}{\sum_t GT_t}. \quad (1)$$

Results for some of the sequences in the dataset are given in Table 2. The variation in accuracy corresponds to the differences in detection quality between cameras. An exception is camera 1 which provides good detection but still causes more frequent errors such as switches and track cancellations in occlusion situations. These effects occur more frequently, because camera 1 views down a hallway and larger persons in the front can occlude multiple persons in the back.

We provide the resulting tracks and evaluation scores for all sequences on the dataset website.

4.3. Person Re-identification

Methods Person re-identification is evaluated in a retrieval scenario. We compute a number of image features to find a query track in the video data. We use $8 \times 8 \times 8$-bin color histograms in HSV color space to describe the color distribution of a person's appearance. The loss of structural information in color histograms is compensated for by also computing Color Structure Descriptors (CSDs) [12, 16]. CSDs are histograms computed by moving a sliding window over the object bounding box and, thus, are able to encode spatial information into the histogram. We use the approach and parameters as suggested in [12]. For textural description of a person's appearance, we apply Gabor filters. Using a filter bank (GFB) of 8 orientations and 5 scales, we generate a vector of filter response strengths by max-pooling. Additionally, we compute Local Binary Pattern histograms (LBP) to encode texture information. We use the 36 rotation invariant values $LBP^{ri}_{8,R}$ as described in [17]. In order to compare two tracks, feature distances are computed for each combination of images in the tracks.

	W-NC	W-C	A-NC	A-C
CH + GBF	0.28	0.21	0.21	0.17
CSD + LBP	0.37	0.24	0.29	0.19
all	0.40	0.27	0.33	0.21

Table 3. MAP values for re-identification within and across cameras, with and without clothing changes. The combination of CSD and LBP performs better than the faster color histograms and Gabor filters. A combination of all features yields the best results.

The feature distances for each image pair are combined using uniform weights. A final track distance is computed by choosing the minimum of all image pair distances. The result of the retrieval task is a list of tracks in the video dataset ranked according to their distance to the query track.

We do not use any segmentation of persons but instead limit the area for feature computation to $[x, y, w, h] = [0.35, 0.2, 0.3, 0.4]$ relative to the bounding box. In order to avoid confusions from possible false-positive tracks, we used ground truth tracks as input for the re-identification.

Results For evaluation of the person re-identification task, we use the Mean Average Precision (MAP) metric. The MAP is computed as the mean of the average precisions (APs) of a set of ranked lists:

$$AP(L) = \frac{1}{\sum_{i=1}^{|L|} m_i} \sum_{i=1}^{|L|} m_i \frac{\sum_{j=1}^{i} m_j}{i} \qquad (2)$$

where L is a ranked list of $|L|$ entries, m_i is 1, if the list entry at position i is a correct result and 0 otherwise.

We evaluate four different re-identification scenarios: within camera (W) and across cameras (A), with and without clothing changes (C,NC respectively). We generate ranked lists using each available track as a query and compute the MAP over all those lists. The resulting accuracies for different feature combinations are given in Table 3. Note that AP is a very strict measure and greatly penalizes incorrect matches at top ranks.

We provide the ranked results for all re-identification experiments as well as evaluation scripts on the dataset website.

4.4. Attribute Classification

Methods For soft-biometric attribute detection, we trained attribute classifiers for each of the attributes described in Section 3. Features used for classification include color histograms in RGB, HSV and Lab color spaces for color attributes, HOG features and LBP features. Descriptors of multiple features are combined into a single feature vector and fed into a classifier. Classifiers were evaluated on each image of a track and the dominant prediction was assigned as the track's attribute decision. As training data we used sequences from the first two days and evaluated the classifiers on data from the third day.

Results We evaluated three different classifiers for attribute recognition: AdaBoost, RBF-Kernel Support Vector Machines (SVM) and Random Forests. The AdaBoost classifier yielded the best results for the majority of attributes. Among the attributes which can be most reliably detected are jacket (0.69), gender (0.66), and upper body color (0.70) while more subtle attributes, such as glasses (0.47) and short hair (0.52), were not successfully detected. We include attribute classification scores into the re-identification task by computing the attribute score differences between two tracks and dividing them by the number of attributes, thus obtaining an average attribute distance for the tracks. This distance is combined with the image feature distances described in the previous section. Although the accuracy of individual attribute classifiers is not high, performance for re-identification across cameras (A-NC) can be improved to a MAP of 0.35.

We provide classification scores for all attributes and tracks on the dataset website.

5. Conclusion

In this work we presented a surveillance video dataset which allows for the evaluation of multiple important computer vision tasks. This includes person detection, tracking, re-identification and soft-biometric attribute recognition. The video data contains many of the challenges occuring in real world surveillance data and features sequences of varying levels of complexity. We performed a number of experiments to generate baseline results for each of the tasks as a reference to compare future results to. We will provide the dataset including images, camera calibrations, evaluation tools and our baseline results on the website.

References

[1] CAVIAR: Context Aware Vision using Image-based Active Recognition. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Boosted Human Re-Identification using Riemannian Manifolds. *Image and Vision Computing*, 2012.

[3] D. Baltieri, R. Vezzani, and R. Cucchiara. 3DPeS: 3D People Dataset for Surveillance and Forensics. In *Joint ACM Workshop on Human Gesture and Behavior Understanding (J-HGBU)*, 2011.

[4] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008.

[5] A. Bialkowski, S. Denman, P. Lucey, S. Sridharan, and C. B. Fookes. A Database for Person Re-Identification in Multi-Camera Surveillance Networks. In *International Conference*

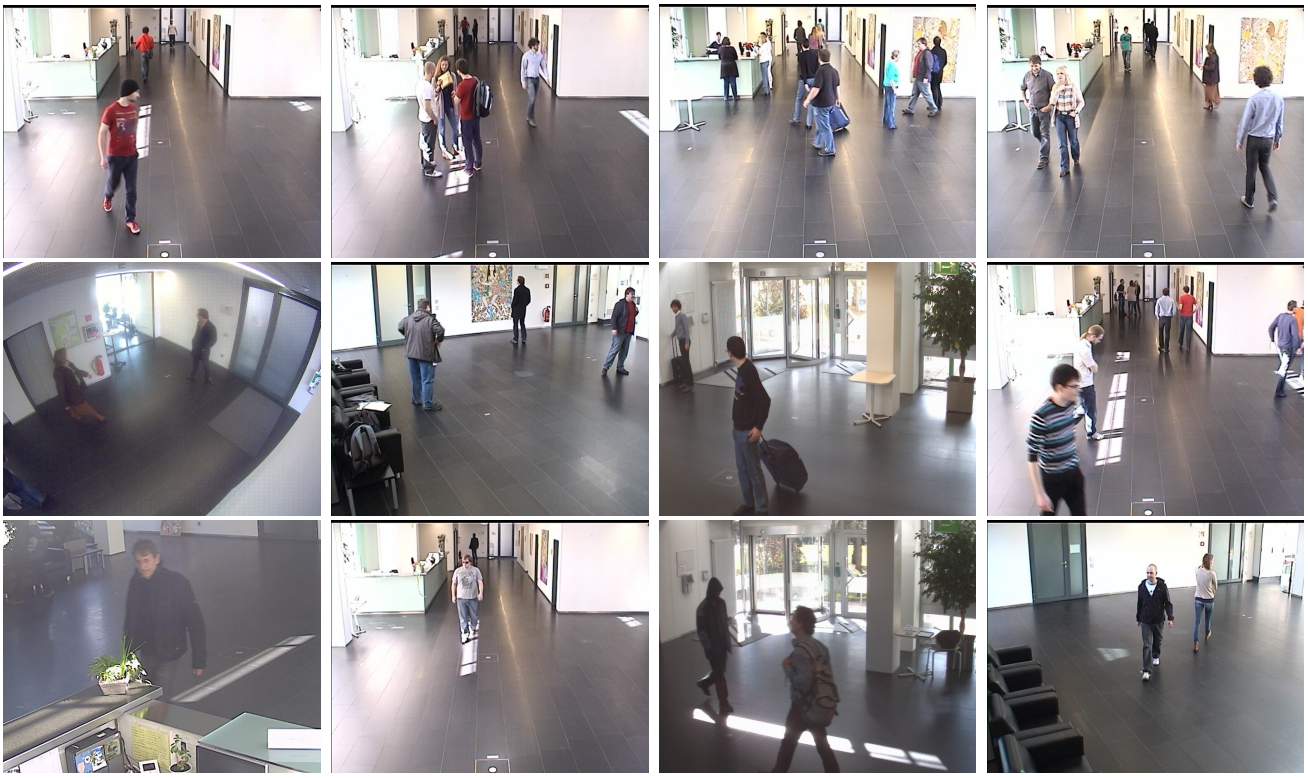

Figure 6. Impressions from the dataset showing different camera views, illumination changes, and varying number of persons in different sequences.

on Digital Image Computing Techniques and Applications (DICTA), 2012.

[6] L. Bourdev and J. Malik. Poselets: Body Part Detectors Trained using 3D Human Pose Annotations. In *International Conference Computer Vision (ICCV)*, 2009.

[7] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom Pictorial Structures for Re-identification. In *British Machine Vision Conference (BMVC)*, 2011.

[8] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[9] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast Feature Pyramids for Object Detection. *Pattern Analysis and Machine Intelligence*, 2014.

[10] A. Ess, B. Leibe, and L. Van Gool. Depth and Appearance for Mobile Scene Analysis. In *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[11] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 2007.

[12] M. Hahnel, D. Klunder, and K.-F. Kraiss. Color and Texture Features for Person Recognition. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2004.

[13] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person Re-Identication by Descriptive and Discriminative Classification. In *Scandinavian Conference on Image Analysis (SCIA)*, 2011.

[14] R. Layne, T. M. Hospedales, and S. Gong. Attributes-based re-identification. In *Person Re-Identification*. Springer, 2014.

[15] A. Nilski. Evaluating Multiple Camera Tracking Systems - The i-LIDS 5th Scenario. In *International Carnahan Conference on Security Technology (ICCST)*, 2008.

[16] J. R. Ohm, L. Cieplinski, H. J. Kim, S. Krishnamacha, B. S. Manjunath, D. S. Messing, and A. Yamada. Color Descriptors. In *Introduction to MPEG-7*. John Wiley & Sons, Inc., 2002.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *Pattern Analysis and Machine Intelligence*, 2002.

[18] A. Schumann, M. Bäuml, and R. Stiefelhagen. Person Tracking-by-Detection with Efficient Selection of Part-Detectors. In *International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2013.

[19] W. R. Schwartz and L. S. Davis. Learning Discriminative Appearance-Based Models Using Partial Least Squares. In *Barazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, 2009.

[20] V. Takala and M. Pietikainen. CMV100: A Dataset for People Tracking and Re-Identification in Sparse Camera Networks. In *International Conference on Pattern Recognition (ICPR)*, 2012.

[21] W.-S. Zheng, S. Gong, and T. Xiang. Person Re-identication by Probabilistic Relative Distance Comparison. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.