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.

Person tracks 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.

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

![Floor plan of the area in which the dataset was recorded. The camera network consists of six IP cameras and one HD camera.](sobisdata.org)
Table 1. Main characteristics of the cameras in the network.

<table>
<thead>
<tr>
<th>Resolution</th>
<th>FPS</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cam1</td>
<td>768×576</td>
<td>20</td>
</tr>
<tr>
<td>Cam2</td>
<td>640×480</td>
<td>20</td>
</tr>
<tr>
<td>Cam3</td>
<td>702×576</td>
<td>10</td>
</tr>
<tr>
<td>Cam4</td>
<td>768×576</td>
<td>10</td>
</tr>
<tr>
<td>Cam5</td>
<td>704×576</td>
<td>15</td>
</tr>
<tr>
<td>Cam6</td>
<td>704×576</td>
<td>15</td>
</tr>
<tr>
<td>Cam7</td>
<td>1440×1080</td>
<td>25</td>
</tr>
</tbody>
</table>

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 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.
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 performance 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

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

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

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

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

<table>
<thead>
<tr>
<th></th>
<th>SS1</th>
<th>SS2</th>
<th>SM1</th>
<th>SG1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera 1</td>
<td>66.2</td>
<td>67.1</td>
<td>58.2</td>
<td>54.3</td>
</tr>
<tr>
<td>Camera 2</td>
<td>65.3</td>
<td>68.0</td>
<td>65.3</td>
<td>61.4</td>
</tr>
<tr>
<td>Camera 5</td>
<td>60.3</td>
<td>61.0</td>
<td>56.4</td>
<td>52.2</td>
</tr>
</tbody>
</table>

was used. As input we used the poselet person detections, due to their high accuracy.

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 (GBF) 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 \(LBP_{ri}^{s,R}\) as described in [17]. In order to compare two tracks, feature distances are computed for each combination of images in the tracks.
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 occurring 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


