# Automated Generation of High-Quality Training Data for Appearance-based Object Models

Stefan Becker, Arno Voelcker, Hilke Kieritz, Wolfgang Hübner, and Michael Arens

Fraunhofer Institute for Optronics, System Technologies, and Image Exploitation IOSB
Gutleuthausstr. 1, 76275 Ettlingen, Germany

## ABSTRACT

Methods for automated person detection and person tracking are essential core components in modern security and surveillance systems. Most state-of-the-art person detectors follow a statistical approach, where prototypical appearances of persons are learned from training samples with known class labels. Selecting appropriate learning samples has a significant impact on the quality of the generated person detectors. For example, training a classifier on a rigid body model using training samples with strong pose variations is in general not effective, irrespective of the classifiers capabilities. Generation of high-quality training data is, apart from performance issues, a very time consuming process, comprising a significant amount of manual work. Furthermore, due to inevitable limitations of freely available training data, corresponding classifiers are not always transferable to a given sensor and are only applicable in a well-defined narrow variety of scenes and camera setups. Semi-supervised learning methods are a commonly used alternative to supervised training, in general requiring only few labeled samples. However, as a drawback semi-supervised methods always include a generative component, which is known to be difficult to learn. Therefore, automated processes for generating training data sets for supervised methods are needed. Such approaches could either help to better adjust classifiers to respective hardware, or serve as a complement to existing data sets.

Towards this end, this paper provides some insights into the quality requirements of automatically generated training data for supervised learning methods. Assuming a static camera, labels are generated based on motion detection by background subtraction with respect to weak constraints on the enclosing bounding box of the motion blobs. Since this labeling method consists of standard components, we illustrate the effectiveness by adapting a person detector to cameras of a sensor network. While varying the training data and keeping the detection framework identical, we derive statements about the sample quality.

**Keywords:** training data, boosting, object detection, video annotation

## 1. INTRODUCTION

Person detection is indispensable for automated video surveillance. An initial task of learning the prototypical appearances of persons consist of collecting labels for training samples, which is usually obtained by labeling manually. This procedure is very tedious and time consuming. Furthermore, the selecting of appropriate samples has a significant impact on the quality of the generated detector. In return, the choice of the learning algorithm and the feature representation influences the requirements on the samples. For example, detectors with a part-based representation model such as the approach of Leibe et al.[1] or Bourdev et al.,[2] which rely on feature selecting during run time, are very flexible concerning a strong intra-class variation in the training data.

In contrast to a representation by its parts, a rigid representation can be used. Here, multi-scale person detection is done with a sliding window approach. For every window location and scale a binary classication based on a fixed feature formation is done. In general, a distinction is made between two major approaches. On

Further author information:
Stefan Becker: E-mail: stefan.becker@iosb.fraunhofer.de
Arno Voelcker: E-mail: arno.voelcker@iosb.fraunhofer.de
Hilke Kieritz: E-mail: hilke.kieritz@iosb.fraunhofer.de
Wolfgang Hübner: E-mail: wolfgang.huebner@iosb.fraunhofer.de
Michael Arens: E-mail: michael.arens@iosb.fraunhofer.de

Emerging Technologies in Security and Defence; and Quantum Security II; and Unmanned Sensor Systems X,
edited by K. L. Lewis, R. C. Hollins, T. J. Merlet, M. T. Gruneisen, M. Dusek, J. G. Rarity, E. M. Carapezza,
Proc. of SPIE Vol. 8899, 889916 · © 2013 SPIE · CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2028714

the one hand, there is the concept of handcrafted features such as histogram of oriented gradients (HOG) (Dalal et al.[3]) using a support vector machine as classifier. On the other hand, features are selected with boosting during the training process to build up an cascaded classifier from simple features such as the approach from Viola et al.[4] or Dollar et al.[5] However, the outstanding characteristic is a fixed feature constellation inside the window either homogeneously distributed over the window or using a sparse feature set with unique positions. Therefore training a classifier on a rigid representation model using training samples with strong pose variations is not effective and can lead to a loss of dominant property and meaningful characteristic.

Despite this seemingly disadvantage, most state-of-the-arts methods follow the sliding window paradigm with such a rigid model. Especially for applications running under real-time conditions, cascade classifiers are currently very popular due to their computational effectiveness. Moreover, with the knowledge about the peculiarity of rigid models, there is the possibility to adjust the training data to representation models instead of adjusting representation models to the data. Divvala et al.[6] showed that the same detector performance can be achieved by either using a deformable part representation with the approach of Felzenszwalb et al.[7] or by a divide-and-conquer strategy resulting in parallel subcategory detectors with a simplified representation model. Instead of handling large intra-class variations by manually clustering the training data, Wu et al.[8] introduced a so called Cluster Boosted Tree to automatically construct tree structured object detectors by dividing the sample space based on features selected by boosting.

In this work, without preferring a specific person detector, we use a detector similar to the work of Dollár et al.[5] based on integral channel features and a classifier trained by AdaBoost. On the one hand, it has been shown to give superior performance on general person datasets (see Dollár et al.[9]) and on the other hand, alternative approaches are expected to have equal or lesser requirements on the training samples quality. Therefore, all automatically labeled training samples which suffice the requirements for this particular learning and detection strategy can also be used for training with almost any other supervised methods.

Contrary to the counterproductive effect of too strong variations, the training data has to capture sufficient variations in order to not overfit to a specific bias in the limited amount of positive samples. Therefore the training samples have to capture sufficient variance in pose, differing clothing etc. Towards this end, general datasets are designed to capture the entire visual appearances of the object class without the presence of any bias. To obtain a larger set of varying positive samples, some approaches address this problem by generating synthetic samples. In Ref.[10] and Ref.,[11] Pishulin et al. start from 3D pose annotations and recover the shape. Then the modified body shape is back-projected into the image. Enzweiler et al.[12] proposed a active appearance models in order to extend the positive training samples. An alternative way of dealing with a dataset bias and for getting a better generalization ability, is for example presented in the work of Khosla et al.[13] They combined several existing datasets to obtain a larger image database by proposing a discriminative framework to learn weights for the individual dataset bias and bias associated with each dataset. These learned weights can be applied to a novel dataset. Vazquez et al.[14] introduced a domain adaptation procedure to better adapt virtual, synthetic samples to real-world images. In contrast to the attempt of creating an unbiased, generic dataset, the question arises how diverse are requirements on a specialized, scene-specific dataset.

The desired scenario of application is a camera network consisting of several stationary cameras which monitor standard indoor scenes (see figure 2). Compared to unstructured environments in such a setup, the amount of possible variations is significantly more restrictive concerning the diversity of person's appearances and background changes. In order to overcome some limitations of a generic dataset or rather a generic detector, there exists several approaches for adapting a detector to a sensor. Here, we want to exemplary mention the approaches of Roth et al.,[15,16] Sternig et al.,[17] Stadler et al.,[18] and Celik et al.[19] The basic idea in Ref.[15] is to use a bootstrap approach to iteratively obtain a better classier for the scene by starting with a moving object classifier which is further used to generate a more complex classifier. In Ref.[16] a concept of classifier grids, in which instead of a sliding window based detection a unique detector for every image location is trained, is introduced. This is done by using a generative representation for the background and for the object class and combines them to get a discriminative model. To adapt to changing environments these classifiers are learned online by a fixed object representation and updating the background representations. The approaches from Sternig et al.[17] and Stadler et al.[18] extend the concept of classifier grids by a co-training strategy to reduce the bias of the initial generative models. Further, various types of context information (e.g. scene knowledge) are integrated to also update the

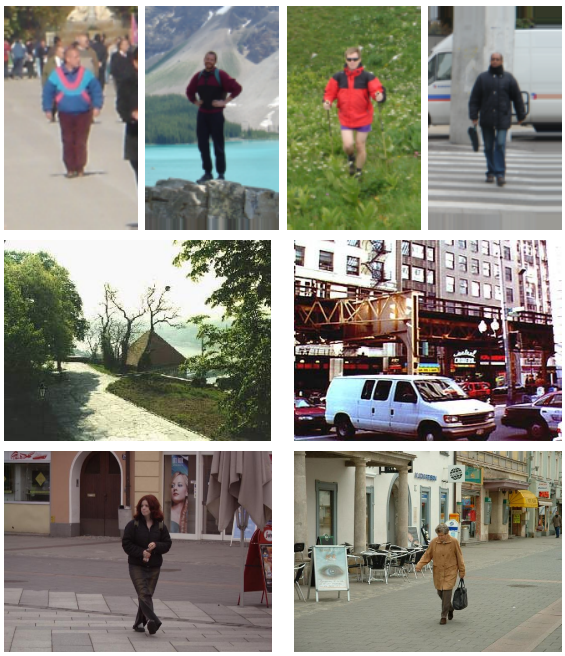


Figure 1. Examples from the INRIA dataset (see Ref.[3]). (Top) Some positive labeled samples from INRIA. Persons are always upright, but with some partial occlusion and a wide range of variations in pose, appearance, clothing, illumination and background. (Middle) Some negative labeled samples from INRIA. (Bottom) Some example images from the test set which consist of overall 288 independent scenarios.

object representation. In Celik et al.,[19] statistical information of moving objects in the scene is collected during a learning phase and then used for filtering out the objects considered as dominant. Filtering is based on the collected statistical information is referred to as their coarse object detector. Afterwards a data refinement via spectral clustering utilizing on a bag of HOG features. Followed by training their fine object detector on the reduced training set with a support vector machine. In Celik et al.,[20] this concept is further expanded by a system of parallel detectors in order to deal with multiple object classes.

Overall, the processing pipeline from Celik et al.[20] is the closest to this work. We also assume a static camera and generate labels for a rigid representation model from motion blobs. Here a central aspect is to investigate how sensitively a rigid representation model with a boosting learning strategy reacts to automatically generated training samples. Our question implies how accurate an unsupervised labeling method should be. Furthermore, we analyze the requirements on a specialized, scene-specific dataset by comparing the benefit from scene-specific samples with samples from a generic dataset.

This article is structured as follows. In section 2 we introduce the datasets and describe the diverse characteristics of the scene-specific and the generic data, followed by a short explanation of the components of the analyzed detector. Our proposed processing pipeline of naive labeling is explained in section 3 including the underlying background subtraction algorithm and post-processing steps. Section 4 depicts the evaluation of the training data. Several detectors are trained on samples generated under varying conditions and compared against each other and to a reference detector which is trained on the generic dataset. In section 5 some conclusions are drawn.

## 2. DATASET & DETECTOR

In this Section the experimental setup is introduced, including the two compared datasets and the analyzed detector.



Figure 2. Examples illustrating the restricted range of variation in a static camera setup. (Top) Some examples from the training set. The person are upright and walk a similar path in the camera field of view. (Middle) Images which show how a the background for a consistent camera setup can change over time. (Bottom) Some examples from the test set. Only images which include at least one person are considered.

### 2.1 Dataset

As reference or rather generic dataset, we use the freely avaible INRIA Pedestrian Dataset.[3] The INRIA dataset consist of images from Navneet Dalal's personal photo collection with a corresponding selection bias. Nevertheless, the persons are usually upright standing, but the samples also capture a wide range of variations in pose, appearance, clothing, illumination, and background. Single samples are almost independent from each other. The INRIA datset is divided in into two subsets: One for training containing 1237 persons and 1218 negative images and the other for testing or evaluation purpose which is made up of 288 images with at least one person in the scene. figure 1 shows some examples from the positive and negative training and from the test set. For each image ground truth data from manual labeling is available in such a way that a person is described by its enclosing bounding box. The labeled samples as well as their mirror reflections serve as training samples. All samples are resized to a common size of $64 \times 128$ pixels which include a padding of approximately 16 pixels on top and at the bottom. The width is set to half the new height while maintaining the center of the sample. The extra padding is added so that features can be retrieved which respond well on the transition of person and background. Negative samples are randomly generated from the 1218 background images, providing a sufficient background variation.

The second scene-specific dataset – called "Distante Biometrie" (DisBio) – captures only one particular, prototypical surveillance scenario. The data was recorded with the "Network Enabled Surveillance and Tracking" (NEST) system (see Monari[21]). It consists of video footage recorded over three days by six cameras with non-overlapping and partially overlapping fields of view. The dataset includes single persons, crowds passing by, and scenes containing complex actions. In order to clearly distinguish between a training and test set, the third day recordings are regarded as test set. The test scenes mainly contain single persons walking in the camera field of view and two persons crossing each other (see figure 2). In this work, only video sequences from the camera with the largest number of different persons are considered, resulting in 32 unique persons for the generation of positive training samples. The video sequences from the used camera (Axis 214 PTZ) were recorded with a

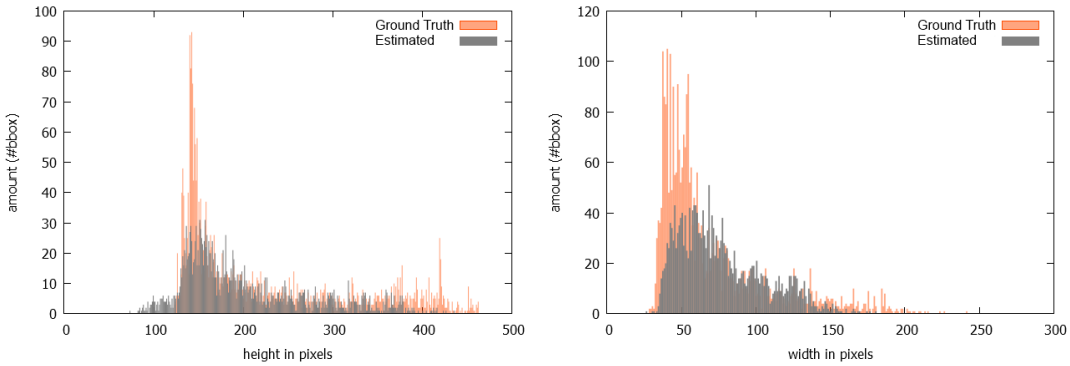



Figure 3. The figures show the width and height distribution between the ground truth and labels estimated via motion detection.

resolution of $768 \times 576$ pixels and a frame rate of 10 frames per second. Ground truth data is available, but only every fifth image is manually annotated by a bounding box. The intermediate frames are interpolated using linear interpolation. figure 2 shows some example image from the Disbio dataset. Those examples are good illustrations on how the range of variations is restricted in a static camera setup.

### 2.2 Detector

Reasons for favouring the used detection framework is motivated in section 1. However, the detector used for evaluation is similar to the work of Dollár et al.[5] They extended the concept of Viola et al.[4] by multiple feature channels in order to generate a set of weak classifiers. The feature channels consist of multiple linear and non-linear transformations of an input image. Here, only a small set of feature channels is used for evaluation. These are, six channels for quantized gradient directions, the gradient magnitude, and the intensity. The classifier is trained by AdaBoost, which combines a set of weak classifiers into a strong classifier. In this work a weak classifier consists of a decision tree of depth two. Each node of a tree uses the sum of one image region on one feature channel to make its decision. The result of the training procedure is an ensemble of decision trees structured as a soft cascade, which can reject an image region after each evaluation of a weak classifier. A final decision is made by thresholding the summed confidence values (see Ref.[22] for details). In order to get a reduced false negative rate, while still be able to reject implausible regions only after a few stages, the threshold is set to the minimum weight of all positive training samples over all weak classifier.

## 3. LABELING

In a scenario with a static camera, the basic idea is to automatically extract labeled data from the monitored scene. All approaches, which collect training samples online face the problem of mislabeling. One contribution of this paper is to address the problem of noisy labels and to quantify their influence. In this section, we compare manual labels with labels generated via motion detection. Manual labels are not necessarily perfect since they are error-prone if various operators mark objects not in a standardized manner (see Ref.[23]). However, the data obtained manually can be considered as the ground truth, which is provided for both datasets (see section 2). For such an application background subtraction is widely used in order to detect moving objects. A good overview of different background subtraction techniques can be found in the survey from Piccardi.[24] Here, we follow the approach of Zivkovic[25]which is based on the work of Stauffer et al.[26] Every image pixel is modeled as a mixture of Gaussian with a per-frame strategy for updating. The classification in foreground and background is made independently for each pixel, therefore the resulting binary mask can benefit from spatial smoothing. We applied a median filtering, followed by a pairwise erode/dilate morphological operation. The contours of the foreground blobs are estimated with the approach of Suzuki et al.[27] taking additionally the merging of nearby scattered blobs into account.The enclosing bounding boxes are filtered with respect to their aspect ratio, only bounding boxes within a range from 1/4 to 1/2 are considered.

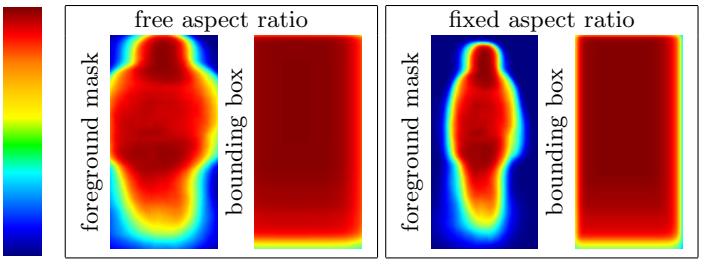



Figure 4. Overlap between ground truth data and estimated labels visualized as heat maps, where blue is associated with a low and red with a high overlap ratio. (Left) Overlap heatmap for the foreground mask and estimated bounding box with a free aspect ratio onto a normalized region. (Right) Overlap heatmap for the foreground mask and estimated bounding box with a fixed aspect ratio onto a normalized region.

Since this processing pipeline of automated labeling is naïve, we compare the resulting labels with the manual labels from a subset of the DisBio training sequences. In this sequences each person enters the camera's field of view from the rear right, walks to the center and then towards the camera. Figure 3 shows the distribution of the width and height of the generated labels compared to the ground truth. In general the width of the enclosing bounding box is stronger influenced by the body pose. Hence, it seems necessary to only consider the main body axis as reference to select positive training samples. Therefore, we keep the height and fit the width according to a fixed aspect ratio from 1/2.

In such a indoor scenario the distributions of the estimated labels appear very similar to the ground truth. However, height and width tends to be underestimated by the naïve labeling, which is clearly visible in the lower level of the height distribution (see figure 3). This results mainly from inaccurate motion segmentation e.g. missing heads, legs, or torso. For quantizing the overlap of the bounding boxes their aspect ratio should be identical. Therefore the data is adjusted that the aspect ratio for each box is 1/2 by altering the width to the half of the height while maintaining each center. These overlaps can be displayed as heat maps which are constructed as the averaged sum of the intersection between the estimated and the ground truth bounding box onto a normalized region with a size of $64 \times 128$ pixels. The heat maps are on one hand established via the boxes themselves and on the other hand by the underlying foreground mask which the bounding boxes refer to. In figure 4 the low overlap ratio at the bottom region stands out most. This is due that the feet are difficult to separate from the background and often missed in the foreground mask. This low overlap could be improved with more knowledge about the scene (e.g. fitting of a ground plane) or stronger constraints on the extracted foreground mask. The highest overlap ratio is around the area of the torso and the head. Clearly visible is the contour of an upright person in the foreground heatmaps with no indication of moving arms, this illustrates that the variations of poses from person in the DisBio training set is limited.

## 4. EXPERIMENTS

Nevertheless, the overall overlap of the manual and the estimated labels could be improved with more scene knowledge, stronger constraints, or an improved background subtraction algorithm, but the resulting labels are always subject to errors, especially in a more challenging camera setup. The question arises how suitable are such samples for training a detector and how strong a person detector benefits from more accurate labels. One way to identify an obvious bias inside a dataset is to calculate the average feature channels. Figure 5 illustrates this exemplary. The image on the right shows the average of the gradient magnitude feature channel of all positive training images of the INRIA dataset. The outline of an upright person is reasonably clear due to accurate labeling and most background clutter is successfully averaged out. The image on the left is calculated on a small amount of specifically selected example images to illustrate a poor alignment.

However, there exists no unique metric to assess the value of single samples within a dataset or to assess the quality of the complete dataset, although this topic attracts currently more attention. Here, we just exemplary mention the work of Torralba et al.,[28] in which the bias between freely available generic datasets is evaluated by a cross validation of the detector performance. Finally, the goal of the detection task is the localization of

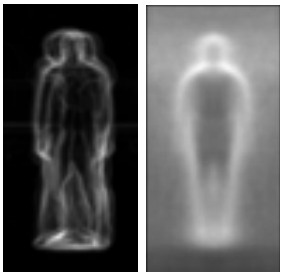


Figure 5. (Left) Average gradient magnitude image over a small selected set of person samples. (Right) The average gradient magnitude image over the positive training examples of the INRIA Dataset.

persons in an image. Hence, by using the same training algorithm on different datasets and keeping the detection framework identical, statements about the data quality can be derived.

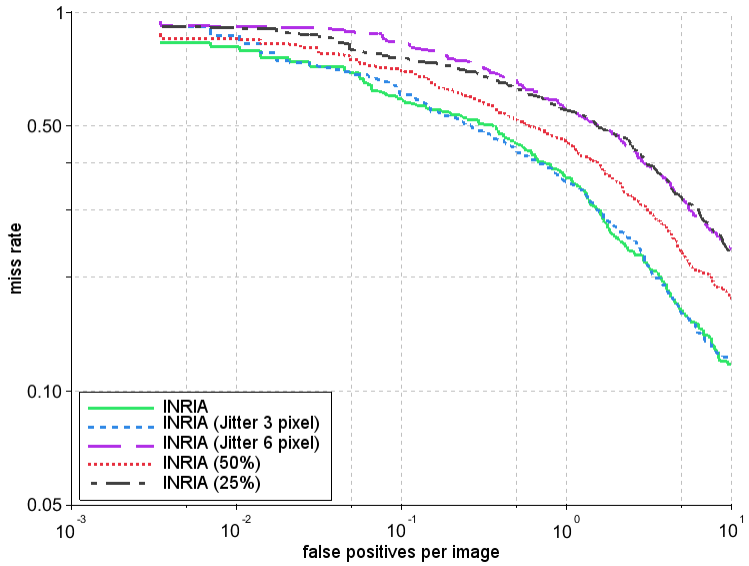



Figure 6. Detection results on the INRIA test set for a detector trained with the unaltered INRIA training set, reduced number of positive samples (25% and 50%), and modified samples with a position jitter of 3 and 6 pixels.

In order to evaluate the detector, we perform a single frame evaluation. Detection is done on multiple scales and a necessary non-maxima suppression for merging nearby detections is applied. For every image a bounding box and a confidence value for each detection is returned. A detected bounding box $bb_{dt}$ and a ground truth bounding box $bb_{gt}$ match if the their overlap exceeds a threshold of 0.4 (PASCAL measure, see Ref.[29]):

$$\frac{area(bb_{dt} \cap bb_{gt})}{area(bb_{dt} \cup bb_{gt})} > 0.4 \tag{1}$$

where $bb_{dt} \cap bb_{gt}$ denotes the intersection of the detected and ground truth bounding boxes and $bb_{dt} \cup bb_{gt}$ their union.

To investigate the influence of poorly labeled training samples, a position noise is added to the positive samples of the generic INRIA datset. Before training, all samples are resized to a size of $64 \times 128$ pixels, keeping the center of the annotated height and adjusting the width. We add a random Gaussian distributed offset to the center position with a standard deviation of 3 and 6 pixels. The performance of these detectors is compared with a detector trained on the standard INRIA dataset and with detectors trained on a reduced number of positive samples. For all following results, we plot the miss rate of the detectors against false positives per image (FPPI). Hence, lower curves indicate better performance. Figure 6 shows the performance of the detectors trained with

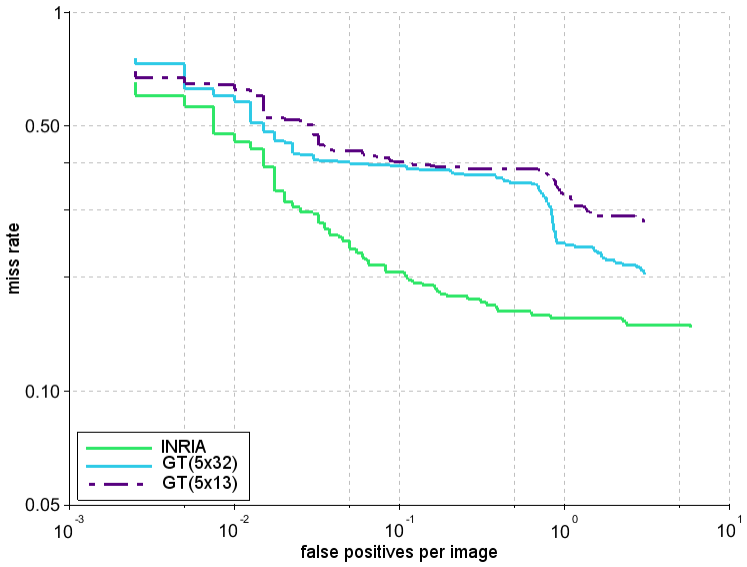



Figure 7. Detection results on the DisBio test set for a detector trained with the INRIA training set and detectors trained only on DisBio including samples from 32 and 13 persons (5 samples and mirror reflections per person ).

the modified sets of positive samples. GT indicates ground truth with the number of samples per person and the number of unique persons (GT(#samples×#persons)). FG denotes the estimated bounding box relying on the foreground mask.

This figure clearly illustrates that the performance drops when the labeling is very poor. In the case of a position jitter with a standard deviation of 6 pixels the detector performance is on a similar level with the detector only trained on 25% of the data. Noticeable is also the fact that a small amount of position jitter has no significant influence on the overall performance. In general, a position offset can be interpreted as adding more variance into the data, which is also needed for boosting in order to better generalize. When training a detector with boosting on the INRIA dataset, the resulting detector is optimized for upright persons (see figure 5). If there is some additional center displacement in the training set, boosting will select larger image regions in the feature channels in order to cope with the additional variance, which can help to detect more pose variation but is less prototypical for a person. However, in order to improve a detector performance by enlarging the number of training samples, the added samples should not exceed a small amount of center displacement to not result in less effective detector. Here, a position jitter of 6 pixels has to be compensated by four times more positive samples.

The last experiment also shows that the increase of training data must be very significant for achieving a significant improvement for a generic detector. In the next experiments, we investigate how effectively a detector can be adapted from scratch to a scene and compare the performance with the generic INRIA detector. Therefore, the amount of scene-specific samples from different persons is iteratively increased in order to achieve a similar performance as a generic detector. The scene-specific sample benefit can be assessed through the reduced number of positive samples which is necessary to achieve comparable results. For this experiment labels are taken from the ground truth. Every person in the DisBio training set that walks in the field of view of the camera, is at least fully visible for 25 frames. In order to avoid too similar training samples, the complete walking sequence per person is equally divided into sub-sequences and from the resulting intervals 5 samples per person are randomly selected. Futher, following the reference detector training, the mirror reflections are also considered. All detectors are trained on the INRIA negative set and with no additional training stage for the collection of hard negative samples by bootstrapping. For evaluation on DisBio sequences, we randomly picked 400 images, where at least one person is present. Figure 7 shows that the reference detector outperforms detectors which only rely on samples from the camera. This result is not surprising despite the fact the DisBio training set only contains a

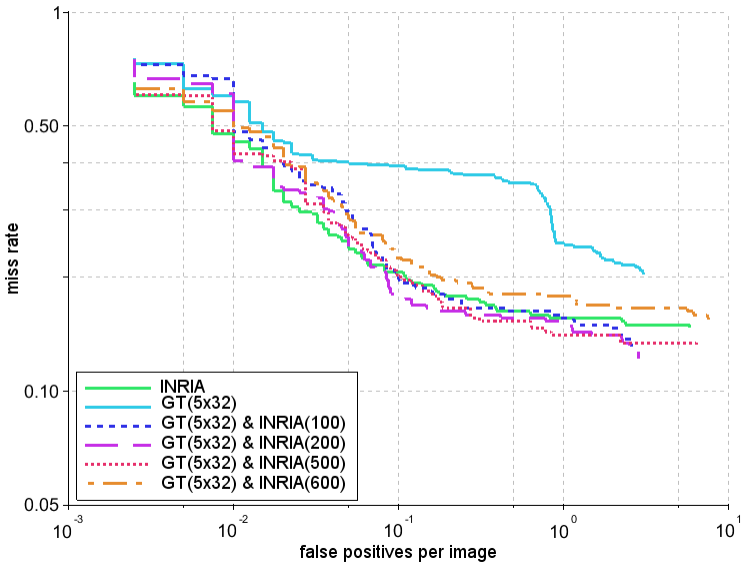



Figure 8. Detection results on the DisBio test set for the reference INRIA detector, and detectors trained on DisBio with an increasing amount of samples from INRIA.

small number of different persons compared to the INRIA training set. Under such conditions, boosting overfits to low variance regions because the number of positive samples is very low. Increasing the number of positive samples by selecting more samples per persons leads to a similar performance, because the samples are strongly correlated and provide insufficient variation to prevent overfitting.

On a generic dataset, performance has a logarithmic dependency on the amount of training data (Torralba et al.[28]). Here, only samples from 13 or 32 persons are present in the training data and hence the limiting factor is the lack of variation. To ensure a better comparability, only a subset of the generic positive samples set can be considered. Due to the fact, that the DisBio camera view angle is very similar to the upright samples from INRIA, the INRIA detector is not only well suited for this camera setup, but the capturing bias of these two datasets might be very low. Hence, in order to get more variation within the data and to ensure a better comparability samples from INRIA are added to the positive DisBio samples. The corresponding performances can be seen in figure 8.

While increasing up to 500 additional INRIA samples, the miss rate drops with increasing FPPI and corresponding detectors outperform the reference detector. However, the detector trained with 600 INRIA samples is not further improving, but performs even worse than detectors trained on lesser samples and the sole INRIA detector. This is due to a gradual domain shift within the combined datasets. First the performance improves because of the additional samples as they provide a higher variation of persons. Then there arises an ambiguity between the two datasets as there are of course differences between both datasets that can not be neglected, e.g. most INRIA samples are taken from ground level, whereas the DisBio camera position is slightly elevated. A similar effect is stated by Torralba et al.[28] As soon as complete datasets with too strong biases are combined, the performance collapses although the amount of available data increases. Nevertheless, this experiment shows how fast a detector can be adapted to a particular scene by video streams. As a result, a detector trained on samples from only 232 different persons outperforms a detector relying on a more than five times higher sample number.

In section 3, we compared the ground truth with the bounding boxes estimated via naïve labeling in a restricted indoor scenario. Figure 9 depicts the resulting detector performances. For each person the set of generated positive samples is equidistantly divided in subsets, then 5 samples per person are randomly selected from different subsets. The resulting detector is compared with a detector trained on the corresponding ground

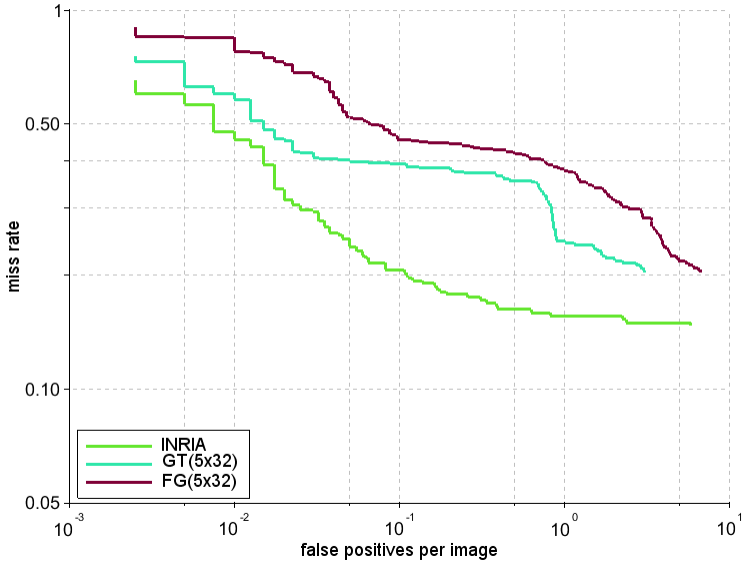



Figure 9. Detection results on the DisBio test set for the reference INRIA detector, and detectors trained on Disbio with ground truth samples and labels generated via motion detection.

truth samples. Although, all samples are estimated from an indoor scene with relative low background dynamic, in which the detection of a single person by motion segmentation is feasible, the automated labeling results in a performance difference in contrast to manual labeling of 10% at a FPPI of $10^{-1}$. This performance loss results from the inaccurate foreground masks and the corresponding center position jitter of the estimated bounding boxes. Nevertheless, in the current processing pipeline potential training samples are only filtered with respect to their aspect ratio and a classifier learned with boosting could be adapted to the scene without manual labeling. No additional data refinement or outlier removal strategy is utilized. In future work, we focus on improving the robustness by including a subsequent processing step that applies a multi-frame registration process to the initially selected bounding boxes.

## 5. CONCLUSION

In this paper, we estimated a reasonable accuracy for centering training samples in order to adjust a classifier trained with boosting to a scene. Within the proposed interval no performance drop could be measured. Moreover this interval can also serve as reference interval for artificially increasing the number of positive samples for a generic dataset. In order to show the effectiveness of scene-specific samples for serving as a complement to existing generic datasets, we ensured a comparability of corresponding detector performances. In our case, a scene-specific detector could outperform a generic detector that was trained with five times more samples. Further, we introduced an automated process for generating training samples via motion segmentation. By relying on standard components and only using weak constraint as outlier removal strategy, manual labeling could be avoided to the cost of a 10% lesser detector performance.

Our initial results showed that the restricted range of variation in surveillance system should be better utilized for overcoming the limitations of generic datasets. In further work, in order to refine the generated training samples a multi-frame registration step is added to the proposed pipeline.

## REFERENCES

1. Leibe, B., Leonardis, A., and Schiele, B., "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision* **77**(1-3), 259–289 (2008).

2. Bourdev, L. and Malik, J., "Poselets: Body part detectors trained using 3d human pose annotations," in [*International Conference on Computer Vision (ICCV)*], (2009).
3. Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], **1**, 886–893 vol. 1 (2005).
4. Viola, P. and Jones, M., "Rapid object detection using a boosted cascade of simple features," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], **1**, I–511–I–518 vol.1 (2001).
5. Dollar, P., Tu, Z., Perona, P., and Belongie, S., "Integral channel features," in [*The British Machine Vision Conference BMVC*], (2009).
6. Santosh K. Divvala, Alexei A. Efros, M. H., "How important are 'deformable parts' in the deformable parts model?," in [*European Conference on Computer Vision Workshops (ECCVW)*], (2012).
7. Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D., "Object detection with discriminatively trained part-based models," *Transactions on Pattern Analysis and Machine Intelligence* **32**, 1627–1645 (2010).
8. Wu, B. and Nevatia, R., "Cluster boosted tree classifier for multi-view, multi-pose object detection.," in [*International Conference on Computer Vision (ICCV)*], 1–8, IEEE (2007).
9. Dollar, P., Wojek, C., Schiele, B., and Perona, P., "Pedestrian detection: An evaluation of the state of the art," *Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **34**(4) (2012).
10. Pishchulin, L., Jain, A., Andriluka, M., Thormaehlen, T., and Schiele, B., "Articulated people detection and pose estimation: Reshaping the future," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], 1–8, IEEE, IEEE, Providence, United States (2012).
11. Pishchulin, L., Jain, A., Wojek, C., Andriluka, M., Thormahlen, T., and Schiele, B., "Learning people detection models from few training samples," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], 1473–1480, IEEE (2011).
12. Enzweiler, M. and Gavrila, D., "A mixed generative-discriminative framework for pedestrian classification," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], 1 –8 (2008).
13. Khosla, A., Zhou, T., Malisiewicz, T., Efros, A., and Torralba, A., "Undoing the damage of dataset bias," in [*European Conference on Computer Vision (ECCV)*], (2012).
14. Vazquez, D., Lopez, A. M., and Ponsa, D., "Unsupervised domain adaptation of virtual and real worlds for pedestrian detection," in [*International Conference on Pattern Recognition (ICPR)*], 3492–3495 (2012).
15. Roth, P., Bischof, H., Skočaj, D., and Leonardis, A., "Object detection with bootstrapped learning," in [*Proc. 10th Computer Vison Winterworkshop*], Hanbury, A. and Bischof, H., eds., 33–42 (2005).
16. Roth, P. M., Sternig, S., Grabner, H., and Bischof, H., "Classifier grids for robust adaptive object detection," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], 2727–2734 (2009).
17. Sternig, S., Roth, P., and Bischof, H., "Learning of scene-specific object detectors by classifier co-grids," in [*Conference on Advanced Video and Signal Based Surveillance (AVSS)*], 408–413 (2010).
18. Stalder, S., Grabner, H., and Gool, L. V., "Exploring context to learn scene specific object detectors," in [*Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*], (2009).
19. Celik, H., Hanjalic, A., Hendriks, E., and Boughorbel, S., "Online training of object detectors from unlabeled surveillance video," in [*Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*], 1–7 (2008).
20. Celik, H., Hanjalic, A., and Hendriks, E. A., "Unsupervised and simultaneous training of multiple object detectors from unlabeled surveillance video," *Computer Vision and Image Understanding* **113**(10), 1076–1094 (2009).
21. Monari, E., *Dynamische Sensorselektion zur auftragsorientierten Objektverfolgung in Kameranetzwerken*, PhD thesis, Karlsruher Instituts für Technologie (KIT) (2011).
22. Bourdev, L. and Brandt, J., "Robust object detection via soft cascade," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], **2**, 236 – 243 (2005).
23. Welinder, P. and Perona, P., "Online crowdsourcing: rating annotators and obtaining cost-effective labels," in [*Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*], (2010).
24. Piccardi, M., "Background subtraction techniques: a review," in [*International Conference on Systems, Man and Cybernetics (SMC)*], 3099–3104 (2004).

25. Zivkovic, Z., "Improved adaptive gausian mixture model for background subtraction," in [*International Conference on Pattern Recognition*], (2004).
26. Stauffer, C. and Grimson, W., "Adaptive background mixture models for real-time tracking," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], (1999).
27. Suzuki, S. and Abe, K., "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing* **30**(1), 32–46 (1985).
28. Torralba, A. and Efros, A. A., "Unbiased look at dataset bias," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], (2011).
29. Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A., "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision* **88**(2), 303–338 (2010).