Feature based person detection beyond the visible spectrum

Kai Jüngling and Michael Arens
FGAN–FOM
76275 Ettlingen, Germany
{juengling,arens}@fom.fgan.de

Abstract

One of the main challenges in computer vision is the automatic detection of specific object classes in images. Recent advances of object detection performance in the visible spectrum encourage the application of these approaches to data beyond the visible spectrum. In this paper, we show the applicability of a well known, local-feature based object detector for the case of people detection in thermal data. We adapt the detector to the special conditions of infrared data and show the specifics relevant for feature based object detection. For that, we employ the SURF feature detector and descriptor that is well suited for infrared data. We evaluate the performance of our adapted object detector in the task of person detection in different real-world scenarios where people occur at multiple scales. Finally, we show how this local-feature based detector can be used to recognize specific object parts, i.e., body parts of detected people.

1. Introduction

Object detection has been subject to extensive research over the past decades. Many of the traditional tracking approaches are based on foreground detection that distinguishes objects from a static background by image subtraction or some more elaborated approach, e.g. [13]. Drawbacks of these systems are the disability to reliably distinguish different object classes and to cope with ego-motion of the recording camera, though extensions in this latter direction have been proposed by [10]. Recent advances in object detection in the visible spectrum ([14], [8], [12], [11]) encourage the use of trainable, class-specific object detectors to detect people in thermal data.

In this paper, we address the problem of detecting people in real-world thermal images. Unlike other approaches, e.g. [4], that assume a static background and a static camera in order to detect objects by simple foreground or motion segmentation, we are able to detect people from a moving camera without assumptions on the environment. In addition to people/scene classification, this object-classification approach is able to distinguish between object classes, which is important since we are not interested in any moving object, like cars (like a motion detection would return it even in an environment with a static camera), but only in people.

Our person detector is build on a state-of-the-art feature based object detector, introduced by Leibe *et al.* in [8]. We make several enhancements to this object detector and adapt it to the specific task of detecting people in thermal data. The approach of local-feature based person detection is specifically applicable in thermal data, because the variation in person appearance is rather limited compared to visible wavelength imagery. In thermal data, people mostly appear as a light region with a contrast to a darker background. Additionally to just detecting persons as a compound, we show how this feature based person detector can be used to classify body-parts of persons, which can be used for further interpretation on the behavior of people. We evaluate the person detector in three thermal image sequences with a total of 2535 person occurrences. These image sequences cover the complete range of difficulties in person detection, i.e., people appearing at different scales, visible from different viewpoint, and occluding each other.

The paper is structured as follows. In section 2 we briefly introduce the basics of the object detection approach and detail the extensions and adaptations we made to it. In section 3, we show how this feature based object detector can be used to classify object-parts, i.e., body-parts of persons in our application. Section 4 elaborates the issues specifically relevant in the application of this object detector to the task of person detection in thermal data. In section 5, we evaluate the person detector in three different infrared image sequences taken from various viewpoints with different cameras.

2. The refined object detection approach

Our person detector is build on a state-of-the-art object detector by Leibe *et al.* [8]. In this section, we briefly describe the training and detection approach and the enhancements we made.

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

2.1. Training

In the training stage, a specific object class is trained on the basis of annotated example images of the desired object category. The training is based on local features that are employed to build an appearance codebook of a specific object category. Leibe *et al*. use a combination of multiple cues to find interest points in the image and then use local Shape Context Descriptors [2] for feature description. Since this combination of multiple interesting point detectors increases computation time, we use the SURF (Speeded Up Robust Features) descriptors described in [1]. This combination of interest point detection and feature description is specifically designed for fast calculation. A specialized GPU implementation presented in [3] attains a performance of 100 frames per second on 640x480 images for feature detection and matching.

The features extracted from the training images on multiple scales are used to build an object category model. For that purpose, features are first clustered in descriptor space to identify reoccurring features that are characteristic for the specific object class. To generalize from the single feature appearance and build a generic, representative object class model, the clusters are represented by the cluster center (in descriptor space). At this point, clusters with too few contributing features are removed from the model since these cannot be expected to be representative for the object category. The feature clusters are the basis for the generation of the Implicit Shape Model (ISM) that describes the spatial configuration of features relative to the object center and is used to vote for object center locations in the detection process. This ISM is built by comparing every training-feature to each representative (cluster center) that was generated in the previous clustering step. If the similarity (euclidean distance in descriptor space) of a feature and the representative is above an assignment threshold, the feature is added to the specific codebook entry. Here, the feature position relative to the object center – the offset – is added to the spatial distribution of the codebook entry with an assignment probability. This probability is based on the similarity and a single feature can contribute to more than one codebook entry (fuzzy assignment).

2.2. Detection

To detect objects of the trained class in images, SURF features are extracted in each input image. These features (the descriptors) are then matched with the codebook, where codebook entries with a match above a threshold t_{sim} are activated and cast votes for object center locations. To allow for fast identification of promising object hypothesis locations, the voting space is divided into a discrete grid in x-, y-, and scale-dimension. Each grid that defines a voting maximum in a local neighborhood is taken to the next step, where voting maxima are refined by mean shift to accurately identify object center locations.

At this point we make two extensions to the work of Leibe *et al.*

First, we do not distribute the vote weights equally over all features and codebook entries, but use the feature similarities to determine the assignment probabilities. By that, features more similar to codebook entries have more influence in object center voting. The assignment probability $p(C_i|f_k)$ of an image feature f_k, codebook entry C_i combination is determined by:

$$p(C_i|f_k) = \frac{\rho(f_k, C_i) + t_{sim}}{t_{sim}}, \tag{1}$$

where $\rho(f_k, C_i)$ is the euclidean distance in descriptor space multiplied by -1. The same distance measure is used for the probability $p(V_{\vec{x}}|C_i)$ of a vote for an object center location $\vec{x}$ when considering a codebook entry C_i. The vote location $\vec{x}$ is determined by the ISM that was learned in training. Here, $\rho(f_k, C_i)$ is the similarity between a codebook representative and a training feature that contributes to the codebook entry.

The overall probability for, and weight of a vote $V_{\vec{x}}$ is:

$$V_{\vec{x}}^w = p(C_i|f_k)p(V_{\vec{x}}|C_i). \tag{2}$$

Second, we solve the problem of the training data dependency. The initial approach by Leibe *et al.* uses all votes that contributed to a maximum to score a hypothesis and to decide which hypotheses are treated as objects and which are discarded. As a result, the voting and thus the hypothesis strength depends on the amount and character of the training data. Features, that frequently occurred in training data result in codebook entries with a large amount of contributing features and thus in a vast of votes for a single object center location with only the evidence of a single image feature. Since a feature-count independent normalization is not possible at this point, this can result in false positive hypotheses with a high strength, generated by just a single or very few false matching image features. To solve this issue, we only count a single vote – the one with the highest similarity of image- and codebook-feature – for an image-feature/hypothesis combination. We hold this approach to be more plausible since a single image feature can only provide evidence for an object hypothesis once.

The score γ of a hypothesis ϕ can thus, without the need for a normalization, directly be inferred by the sum of weights of all I contributing votes:

$$\gamma_\phi = \sum_{i=1}^{I} V_i^w \tag{3}$$

Certainly, this score is furthermore divided by the volume of the scale-adaptive search kernel (see [8] for details), which is necessary because objects at higher scales can be expected to generate much more features than those on lower scales. Additionally, this enhancement provides us with an unambiguousness regarding the training-feature that created the involvement of a specific image feature in a certain hypothesis. This allows for decisive inference from a feature that contributed to an object hypothesis back to the training data. This is important for the classification of body-parts which is described in detail in section 3.

The result of the detection step is a set of object hypotheses Φ, each annotated with a score γ_ϕ. This score is subject to a further threshold application. All object hypotheses below that threshold are removed from the detection set Φ.

3. Body-part classification

As mentioned in section 2.2, our enhancements provide us with an unambiguousness regarding the training-feature that created a specific vote. This unambiguous inference together with an object-part annotation of the training data, i.e., a body-part annotation of persons, allows for object-part classification. The training data body-part annotation can directly be used to annotate training-features found on body-parts with semantic body-part-identifiers. This annotation is added to the codebook entries for the features that can be associated with certain body-parts. The object hypotheses resulting from detection consist of a number of votes. These were generated by specific entries (that refer to training features) in certain codebook entries that were activated by image features. Using the annotation of these entries, we are now able to infer the semantics of (some) image features that contribute to an object hypothesis.

This body-part classification approach has the weakness that the similarity between an image feature and the training feature is calculated only indirectly by the similarity between the (generalized) codebook representative and the image feature (see equation 1). This means that a feature that is annotated with a body-part and resides in a specific codebook entry could contribute to a person hypothesis because the similarity between an image feature and the codebook representative is high enough (this similarity constraint is rather weak since we want to activate all similar structures for detection) but the image feature does in fact not represent the annotated body-part.

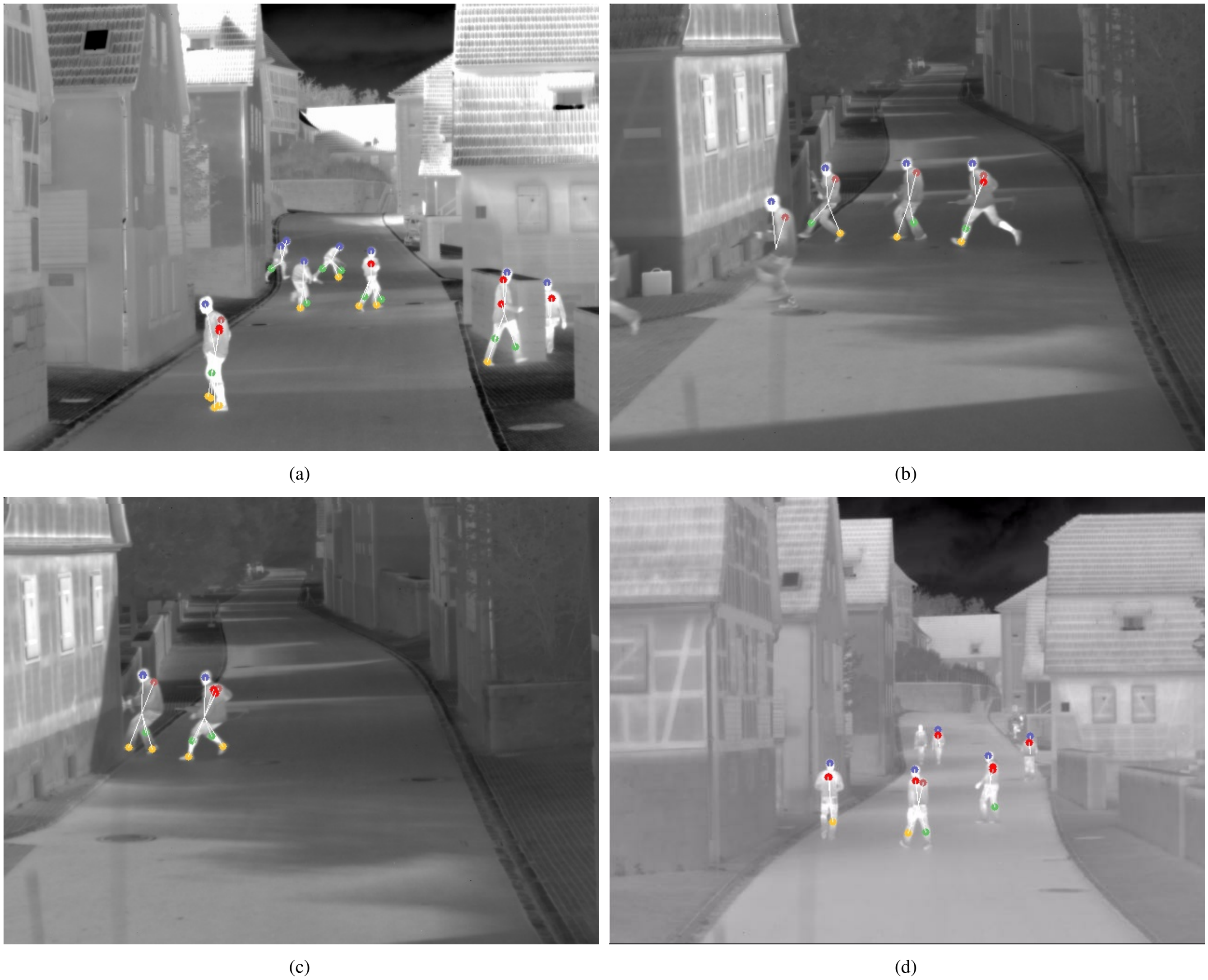
For this reason, we decided to launch another classification level that includes stronger constraints on feature similarity and introduces a body-part-specific appearance generalization. Following that, we generate body-part templates for every body-part class found in the training data. I.e., we pick all features annotated with “foot” from the training data. The descriptors of these features are then clustered in descriptor space to generate body-part templates. The presets on descriptor similarity applied here are stricter than those used in codebook training. This is because we rather want to generate an exact representation than generalize too much from different appearances of certain body-parts. The clustering results in a number of disjoint clusters that represent body-parts. The number of items in a cluster is a measure for how generic it represents a body-part. The more often a certain appearance of a body-part has been seen in training-data, the more generic this appearance is. Since the goal is to create an exact (strong similarity in clustering) and generic (repeatability of features) representation, we remove clusters with too few associated features. The remaining clusters are represented by their cluster center and constitute the templates. These templates can now be used to verify the body-part classification of stage one by directly comparing the feature descriptors of a classified image feature with all templates of the same body-part class. If a strong similarity constraint is met for any of the templates, the classification is considered correct. Otherwise, the image feature annotation is removed.

This multi-level body-part classification is capable of improving the classification robustness by integrating generalized knowledge of body-part appearance which cannot be covered by the codebook generalization itself (since the generalization at that point lacks the body-part information and generalizes only based on appearance). It is not intended to replace the pre-classification by inference completely since this would discard the useful spatial information that is provided by the codebook (that specific body-parts can only be at specific positions – this is covered by the ISM model).

Example results of the body-part classification are shown in figure 1. Here, the relevant body-part categories are: head(blue), torso(red), shoulder(light red), leg(green), and foot(yellow). We see that we are not able to detect every relevant body-part in any case, but the hints can be used – especially when considering temporal development – to build a detailed model of a person which can be the starting point for further interpretation on person behavior. The body-part trajectories acquired in the temporal consideration can then be used to classify the behavior of a person based on the observed articulation. For this, the trajectories can be considered in the reference system of the person itself and a classification, e.g. in “running” or “walking” is possible by just observing the trajectories of certain limbs like feet. (Such an interpretation is actually ongoing work at our lab). The feature-based detection approach is particularly suited for that because it works despite camera motion and changing environment conditions.

4. Detection specifics in infrared data

As mentioned in section 2.1, we use SURF instead of a combination of multiple interesting point detectors and shape descriptors. Despite the faster calculation, these features are particularly suited to distinguish between light



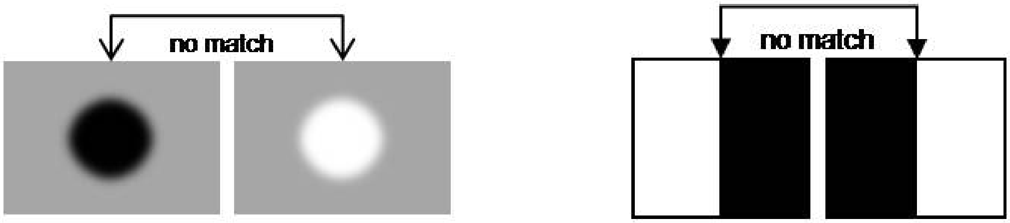
(a) (b) (c) (d)

Figure 1. Example body-part classification results of detected persons. Relevant body-part classes are: head(blue), torso(red), shoulder(light red), leg(green), and foot(yellow).

(foreground) and dark (background) patterns. This is due to the gradient orientations that are used to describe a certain image region. Practically this means – as shown in figure 2(a) – that the descriptor of a light region on dark background does not match a dark region on light background and vice versa. This characteristic is specifically useful in the task of person detection in thermal data because persons typically have a significant light appearance here.

To capture the spatial distribution of the features more accurately, we use the upright version of SURF, the U-SURF features. These are not rotational invariant which increases the distinctiveness. This enables us, as shown in figure 2(b), to distinguish, e.g. features found on the right and left shoulder of persons (left and right refers to the appearance in the image, not right and left from persons view). Especially in our application, where a spatial distribution of features is learned (ISM), this is important because, e.g. a match of the right shoulder in training data with a left shoulder in input data would result in a vote for the wrong object center location.

The object detector used here inherently comprises a bunch of parameters that are relevant for the object detection performance. These parameters comprise settings of the feature extraction, parameters of the training process and various settings of the detection itself. When detecting people in thermal images, especially the settings for the similarity threshold (see section 2.1 and 2.2) are important. In contrast to the application to visible wavelength images, these should be set rather low, which means we require a high similarity in a codebook entry and especially between image feature and codebook entry in detection. This is necessary due to the very similar appearance of different object



(a)

(b)

Figure 2. (a) By using gradient orientation, SURF features of light regions on dark background do not match dark regions on light background.(b) USURF is able to distinguish these kinds of patterns.

parts in thermal data. Since here, the contrast within a person is very low, most of the features found on people refer to parts which have a high contrast with darker background. These most often appear as a light region on dark background which makes the distinctiveness accomplished by U-SURF and the strong similarity requirements inevitable.

5. Experimental Results

5.1. Training data

A crucial point in the performance of a trainable object detector is the choice of the training data used to generate a model of the desired object category. Our person detector is trained with a set of 30 training images taken from an image sequence that was acquired from a moving camera in urban terrain with a resolution of 640x480. The set contains 8 different persons appearing at multiple scales and viewpoints. Two sample images of this set are shown in figure 3. The persons are annotated with a reference segmentation which is used to choose relevant features to train the person detector. Additionally, we annotate the training features with body-part identifiers when this is adequate (when a feature visually refers to a certain body-part). Example results for the body part detection are shown in figure 1. All detection results shown hereafter do not contain any of the persons that appear in training data.

5.2. Person detection

To show the operationality of the detection approach in infrared images, we evaluate the performance in three different image sequences, taken from different cameras under varying environmental conditions. For evaluation, all persons whose head or half of the body is visible, are annotated with bounding boxes.

To assess the detection performance, we use the performance measure

$$recall = \frac{|true\ positives|}{|ground\ truth\ objects|} \tag{4}$$

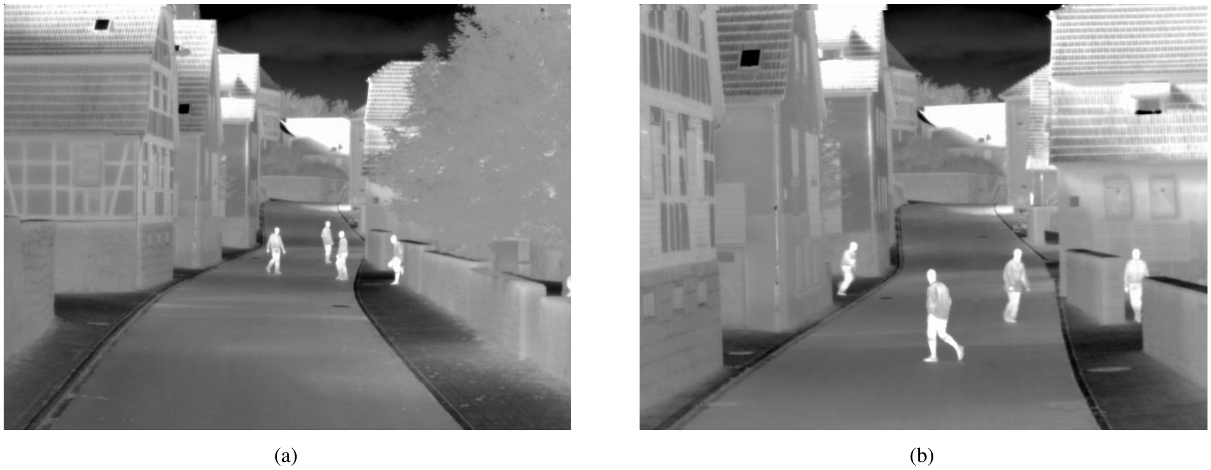
following [9]. To determine whether an object hypothesis is a true- or a false positive, we use two different criteria. The *inside bounding box* criterion assesses an object hypothesis as true-positive if its center is located inside the ground truth bounding box. Only a single hypothesis is counted per ground-truth object, all other hypotheses in the same box are counted as false positive. The *overlapping* criterion assesses object-hypotheses using the ground-truth and hypotheses bounding boxes. The overlap between those is calculated by the Jaccard-Index [7] (compare intersection-over-union criterion [6]):

$$overlap = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}. \tag{5}$$

The first criterion is deliberately used to account for inaccuracies in bounding boxes in the ground truth data and to assess the detection performance independently of its accuracy. Specifically in our case, where the bounding box is defined by the minimal box that contains all features that voted for a hypothesis, a hypothesis that only contains the upper body of a person would be counted as false positive using the overlapping criterion, even if all body-parts of the upper body are correctly found. To depict the accuracy of detections, we use the overlapping criterion which is evaluated for different overlap demands.

The first image sequence contains a total of 301 person occurrences, appearing at roughly the same scale. People run from right to left in the camera's field of view with partial person-person overlapping. We evaluate the sequence using the recall criterion and the false positives per image. The recall is shown as a function of false positives per image as used in various object detector evaluations. To assess the accuracy of the detection we evaluate with different requirements of overlapping. The results for the different evaluation criteria (OLx: Bounding box overlap with a minimum overlap of $x\%$; BBI: Inside bounding box) are shown in figure 5(a). The curves are generated by running the object detector with different parameter settings on the same image sequence. Example detections for this image sequence are shown in figure 4 (a)-(c).

The second image sequence is from OTCBVS dataset [5]



(a)

(b)

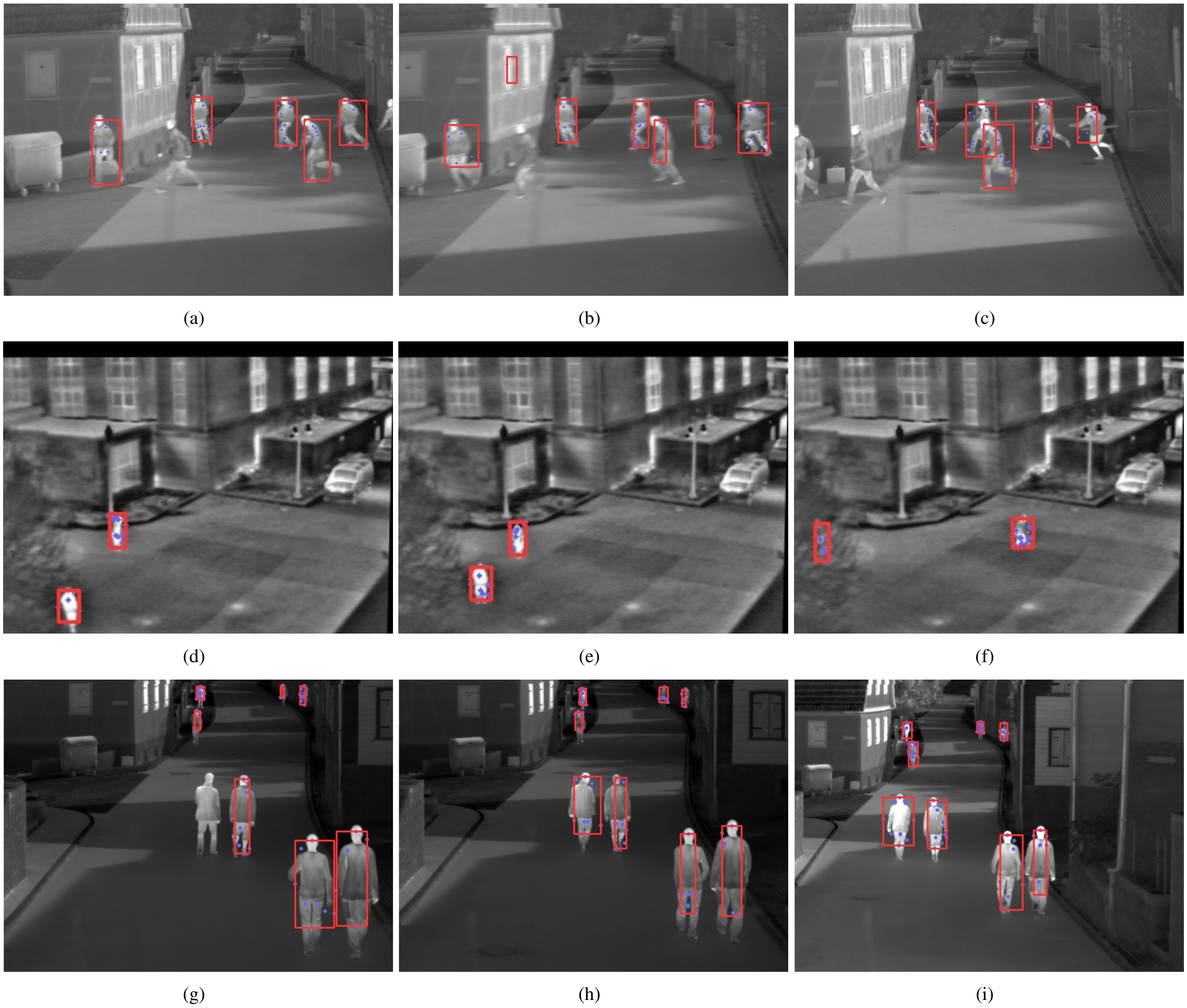
Figure 3. Example images of the training set. The training set comprises 30 images with 8 different persons.

with 763 person occurrences. Here, a scene is observed by a static camera with a high-angle shot. Two persons appearing at a low scale move in the scene without any occlusions. As we see in 5(c), the detection performance is very similar for all false positive rates. Here, we nearly detect all person occurrences in the image at low false positive rates. The results do not improve significantly with other parameters that allow person detections with lower similarity demands and result in more false positives. It is worth mentioning that the detector was trained on persons the appearance of which was not even close to the ones visible in this image sequence. Both, viewpoint and scale of the persons have changed completely between training and input data. Note that the buckling in the curves of bounding box overlap can result from parameter adjustment in allowed feature similarity for detection. Activating more image features for detection can result in more false positive hypotheses and in additional inaccuracies in the bounding box and thus in less true-positives regarding the overlap criterion. The detailed trend of false positives per image and recall for different overlap demands in figure 5(d) shows that the detection performance itself is very good. The accuracy is rather poor compared to the detection performance but still has a recall of above 0.7 with a 50% bounding-box overlap demand. With increasing overlap demand, the detection rate decreases and the false positives increase. As we can see from the development of the curves, this is just due to inaccuracy and not due to "real" false positives generated from background or other objects. Example detections for this image sequence are shown in figure 4 (d)-(f).

The third image sequence has been taken in urban terrain from a camera installed on a moving vehicle. This image sequence, with a total of 1471 person occurrences in it, is the most challenging because a single image contains persons at various scales and the moving paths of persons cross, which leads to strong occlusions. From the example result images in figure 4 (g)-(i), we see that some persons in the background occupy only few image pixels while other persons in the foreground take a significant portion of the whole image. Unlike one could expect, the fact that people are moving parallel to the camera is not very advantageous for the object detector because the persons limbs are not visible very well from this viewpoint. The results of this image sequence are shown in figure 5(b). We see, that the *inside bounding box* criterion performs well and has a recall of more than 0.9 with less than 1.5 false positive/image. When applying the bounding box overlap criterion, the performance drops significantly – stronger than in image sequence one and two. Especially the 50% overlap criterion only reaches a recall of 0.5 with more than 5 false positives/image. This rapid performance degradation is mainly due to the inaccuracies in the bounding boxes of persons appearing at higher scales. This is also visible in the example detections in figure 4 (g)-(i). Here, people in the scene background are most often detected accurately while persons close to the camera are detected rather imprecisely.

6. Conclusion

In this paper we presented a feature based person detection approach with integrated body-part classification and its application to thermal data. We evaluated the person detector in three thermal image sequences with different challenges for person detection. The evaluation results of the three test sequences show that the detector performs well in detecting people per se, but is rather imprecise in terms of bounding boxes. This is inherently due to the feature based approach, that does not account for an object segmentation but determines the bounding box based on local features. The accuracy can be improved by using the segmentation approach of [8] where the backprojection of fea-


(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure 4. Example detections of all 3 evaluation sets. Sequence1:(a)-(c), Sequence 2:(d)-(f), Sequence 3:(g)-(i). Blue points indicate features that generate the hypothesis marked with the red bounding box.

tures is used to obtain a detailed figure/ground segmentation. We discarded the application of this here, because our work aims at doing further processing and interpretations of object-detections on the feature/body-part level.

References

[1] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *Proc. 9th European Conference on Computer Vision*, pages 404–417, Graz, Austria, May 2006.

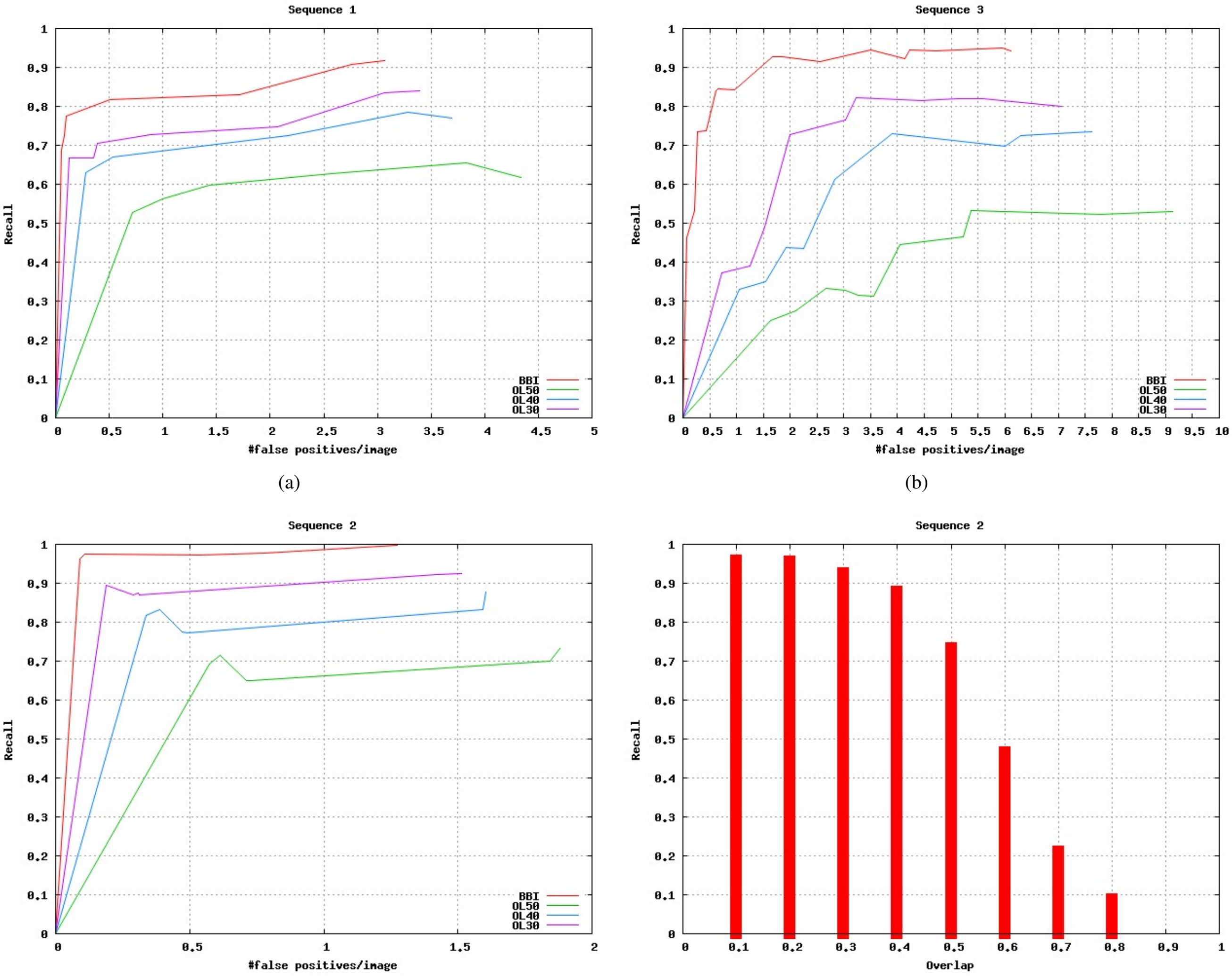
[2] S. Belongie, J. Malik, and J. Puchiza. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, 2002.

[3] N. Cornelis and L. V. Gool. Fast scale invariant feature detection and matching on programmable graphics hardware. In *Computer Vision and Pattern Recognition Workshops*, pages 1–8, Anchorage, USA, June 2008.

[4] J. Davis and V. Sharma. Robust background-subtraction for person detection in thermal imagery. In *Proc. Conference on Computer Vision and Pattern Recognition Workshop*, page 128, 2004.

[5] J. Davis and V. Sharma. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2–3):162–182, 2007.

[6] M. Everingham et al. The 2005 pascal visual object class challenge. In *In Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, 2006.



(a)

(b)

(c)

(d)

Figure 5. Recall/false positive curves for evaluation set 1:(a), 2:(c), 3:(b). Each chart contains four curves that refer to the different evaluation criteria. BBI: Inside bounding box criterion. OL30/40/50: Bounding box overlap criterion with 30, 40 and 50% overlap demand. Figure(d):Trend of detection performance of sequence 2 with a single parameter set using different bounding box overlap demands (displayed on the x-axis in 10% steps).

[7] P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise de Sciences Naturelles*, 4(3):223–370, 1908.

[8] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision*, 77:259–289, 2008.

[9] B. Leibe, K. Schindler, N. Cornelis, and L. V. Gool. Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(10):1683–1698, 2008.

[10] Y. Ren, C.-S. Chua, and Y.-K. Ho. Statistical background modeling for non-stationary camera. *Pattern Recognition Letters*, 24(1–3):183–196, 2003.

[11] E. Seemann, M. Fritz, and B. Schiele. Towards robust pedestrian detection in crowded image sequences. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Mineapolis, USA, June 2007.

[12] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1582–1588, New York, USA, June 2006.

[13] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 246–252, Ft. Collins, CO, USA, June 1999.

[14] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 511–518, Kauai, HI, USA, December 2001.