Augmented Vision and Reality 1

Riad Hammoud
Guoliang Fan
Robert W. McMillan
Katsushi Ikeuchi *Editors*

Machine Vision Beyond Visible Spectrum

Springer

Augmented Vision and Reality

Volume 1

For further volumes:
http://www.springer.com/series/8612

Riad Hammoud · Guoliang Fan
Robert W. McMillan · Katsushi Ikeuchi
Editors

Machine Vision Beyond Visible Spectrum

Springer

Editors

Dr. Riad Hammoud
DynaVox Mayer-Johnson
Wharton Street 2100
Pittsburgh PA 15203
USA
e-mail:
Riad.Hammoud@dynavoxtech.com

Robert W. McMillan
U.S. Army Space and Missile Defense
Command
PO Box 1500
Huntsville AB 35807-3801
USA
e-mail: bob.mcmillan@us.army.mil

Guoliang Fan
School of Electrical and Computer
Engineering
Oklahoma State University
202 Engineering South
Stillwater OK
USA
e-mail: guoliang.fan@oksate.edu

Dr. Katsushi Ikeuchi
Institute of Industrial Science
University of Tokyo
Komaba 4-6-1
Meguro-ku, Tokyo
153-8505 Japan
e-mail: ki@cvl.iis.u-tokyo.ac.jp

ISSN 2190-5916
e-ISSN 2190-5924
ISBN 978-3-642-11567-7
e-ISBN 978-3-642-11568-4
DOI 10.1007/978-3-642-11568-4
Springer Heidelberg Dordrecht London New York

Cover design: eStudio Calamar, Berlin/Figueres

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The genesis of this book on “Machine Vision Beyond the Visible Spectrum” is the successful series of seven workshops on Object Tracking and Classification Beyond the Visible Spectrum (OTCBVS) held as part of the IEEE annual Conference on Computer Vision and Pattern Recognition (CVPR) from 2004 through 2010. Machine Vision Beyond the Visible Spectrum requires processing data from many different types of sensors, including visible, infrared, far infrared, millimeter wave, microwave, radar, and synthetic aperture radar sensors. It involves the creation of new and innovative approaches to the fields of signal processing and artificial intelligence. It is a fertile area for growth in both analysis and experimentation and includes both civilian and military applications. The availability of ever improving computer resources and continuing improvement in sensor performance has given great impetus to this field of research. The dynamics of technology “push” and “pull” in this field of endeavor have resulted from increasing demand from potential users of this technology including both military and civilian entities as well as needs arising from the growing field of homeland security. Military applications in target detection, tracking, discrimination, and classification are obvious. In addition to this obvious use, Machine Vision Beyond the Visible Spectrum is the basis for meeting numerous security needs that arise in homeland security and industrial scenarios. A wide variety of problems in environmental science are potentially solved by Machine Vision, including drug detection, crop health monitoring, and assessment of the effects of climate change.

This book contains 10 chapters, broadly covering the subfields of *Tracking and Recognition in the Infrared, Multi-Sensor Fusion and Smart Sensors*, and *Hyperspectral Image Analysis*. Each chapter is written by recognized experts in the field of machine vision, and represents the very best of the latest advancements in this dynamic and relevant field.

The first chapter entitled “Local Feature Based Person Detection and Tracking Beyond the Visible Spectrum”, by Kai Jüngling and Michael Arens of FGAN-FOM in Germany, addresses the very relevant topic of person detection and tracking in infrared image sequences. The viability of this approach is demonstrated by person detection and tracking in several real world scenarios.

"Appearance Learning by Adaptive Kalman Filters for Robust Infrared Tracking" by Xin Fan, Vijay Venkataraman and Joseph Havlicek of Oklahoma State University, Dalian Institute of Technology, and The University of Oklahoma, casts the tracking problem in a co-inference framework, where both adaptive Kalman filtering and particle filtering are integrated to learn target appearance and to estimate target kinematics in a sequential manner. Experiments show that this approach outperforms traditional approaches with near-super-pixel tracking accuracy and robust handling of occlusions. Chapter 3, "3D Model-Driven Vehicle Matching and Recognition", by Tingbo Hou, Sen Wang, and Hong Qin of Stony Brook University, treats the difficult and universal problem of vehicle recognition in different image poses under various conditions of illumination and occlusion. A compact set of 3D models is used to represent basic vehicle types, and pose transformations are estimated by using approximated vehicle models that can effectively match objects under large viewpoint changes and partial occlusions. Experimental results demonstrate the efficacy of this approach with the potential for extending these methods to other types of objects. The title of Chap. 4 is "Pattern Recognition and Tracking in Infrared Imagery" by Mohammad Alam of the University of South Alabama. This chapter discusses several target detection and tracking algorithms and compares the results obtained to real infrared imagery to verify the effectiveness of these algorithms for target detection and tracking. Chapter 5 describes "A Bayesian Method for Infrared Face Recognition" by Tarek Elguebaly and Nizar Bouguila of Concordia University. It addresses the difficult problem of face recognition under varying illumination conditions and proposes an efficient Bayesian unsupervised algorithm for infrared face recognition, based on the Generalized Gaussian Mixture Model.

Chapter 6, entitled "Fusion of a Camera and Laser Range Sensor for Vehicle Recognition", by Shirmila Mohottala, Shintaro Ono, Masataka Kagesawa, and Katsushi Ikeuchi of the University of Tokyo, combines the spatial localization capability of the laser sensor with the discrimination capability of the imaging system. Experiments with this combination give a detection rate of 100 percent and a vehicle type classification rate of 95 percent. Chapter 7 presents "A System Approach to Adaptive Multimodal Sensor Designs", by Tao Wang, Zhigang Zhu, Robert S. Krzaczek and Harvey E Rhody of the City College of New York, based on the integration of tools for the physics-based simulation of complex scenes and targets, sensor modeling, and multimodal data exploitation. The result of this work is an optimized design for the peripheral-fovea structure and a system model for developing sensor systems that can be developed within a simulation context.

Chapter 8, entitled "Statistical Affine Invariant Hyperspectral Texture Descriptors Based on Harmonic Analysis" by Pattaraporn Khuwuthyakorn, Antonio Robles-Kelly, and Jun Zhou of the Cooperative Research Centre for National Plant Biosecurity in Australia, focuses on the problem of recovering a hyperspectral image descriptor based on harmonic analysis. This chapter illustrates the robustness of these descriptors to affine transformations and shows their utility for purposes of recognition. "Tracking and ID via Object Reflectance Using a Hyperspectral Video Camera" is the title of Chap. 9. This chapter is authored by

Hien Nguyen, Amit Banerjee, Phil Burlina, and Rama Chellappa of the University of Maryland and focuses on the problem of tracking objects through challenging conditions, such as rapid illumination and pose changes, occlusions, and in the presence of confusers. This chapter demonstrates that the near-IR spectra of human skin can be used to distinguish different people in a video sequence. The final chapter of this book, “Moving Object Detection and Tracking in Forward Looking Aerial Imagery”, by Subhabrata Bhattacharya, Imran Saleemi, Haroon Idrees, and Mubarak Shah of the University of Central Florida, discusses the challenges of automating surveillance and reconnaissance tasks for infrared visual data obtained from aerial platforms. This chapter gives an overview of these problems and the associated limitations of some of the conventional techniques typically employed for these applications.

Although the inspiration for this book was the OTCVBS workshop series, the subtopics and chapters contained herein are based on new concepts and new applications of proven results, and not necessarily limited to IEEE OTCBVS workshop series materials. The authors of the various chapters in this book were carefully chosen from among practicing application-oriented research scientists and engineers. All authors work with the problems of machine vision or related technology on a daily basis, and all are internationally recognized as technical experts in the fields addressed by their chapters.

It is the profound wish of the editors and authors of this book that it will be of some use to practicing scientists and engineers in the field of machine vision as they endeavor to improve the systems on which so many of us rely for safety and security.

June 2010

Riad Hammoud
Guoliang Fan
Robert W. McMillan
Katsushi Ikeuchi

Contents

Part I Tracking and Recognition in Infrared

Part II Multi-Sensor Fusion and Smart Sensors

Part III Hyperspectral Image Analysis

Part I
Tracking and Recognition in Infrared

Local Feature Based Person Detection and Tracking Beyond the Visible Spectrum

Kai Jüngling and Michael Arens

Abstract One challenging field in computer vision is the automatic detection and tracking of objects in image sequences. Promising performance of local features and local feature based object detection approaches in the visible spectrum encourage the application of the same principles to data beyond the visible spectrum. Since these dedicated object detectors neither make assumptions on a static background nor a stationary camera, it is reasonable to use these object detectors as a basis for tracking tasks as well. In this work, we address the two tasks of object detection and tracking and introduce an integrated approach to both challenges that combines bottom-up tracking-by-detection techniques with top-down model based strategies on the level of local features. By this combination of detection and tracking in a single framework, we achieve (i) automatic identity preservation in tracking, (ii) a stabilization of object detection, (iii) a reduction of false alarms by automatic verification of tracking results in every step and (iv) tracking through short term occlusions without additional treatment of these situations. Since our tracking approach is solely based on local features it works independently of underlying video-data specifics like color information—making it applicable to both, visible and infrared data. Since the object detector is trainable and the tracking methodology does not make any assumptions on object class specifics, the overall approach is general applicable for any object class. We apply our approach to the task of person detection and tracking in infrared image sequences. For this case we show that our local feature based approach inherently

K. Jüngling (✉) · M. Arens
Fraunhofer IOSB, Gutleuthausstrasse 1 76275 Ettlingen, Germany
e-mail: kai.juengling@iosb.fraunhofer.de

M. Arens
e-mail: michael.arens@iosb.fraunhofer.de

R. Hammoud et al. (eds.), *Machine Vision Beyond Visible Spectrum*,
Augmented Vision and Reality, 1, DOI: 10.1007/978-3-642-11568-4_1,

allows for object component classification, i.e., body part detection. To show the usability of our approach, we evaluate the performance of both, person detection and tracking in different real world scenarios, including urban scenarios where the camera is mounted on a moving vehicle.

Keywords Person detection · Person tracking · Visual surveillance · SURF

1 Introduction

Object, and specifically person or pedestrian detection and tracking has been subject to extensive research over the past decades. The application areas for this are vast and reach from video surveillance, thread assessment in military applications, driver assistance to human computer interaction. An extensive review of the whole field of pedestrian detection and tracking is beyond the scope of this paper and can be found in [11, 18, 40]. We will indicate, however, some representative work for each of what we think to be escalating levels of difficulty: (i) person detection, (ii) person tracking and (iii) person detection and tracking from moving cameras.

Early systems in person centered computer vision applications mainly focused on surveillance tasks with stationary cameras. Here, full systems like [16, 37] built on foreground detection methods that model the static background and detect persons as foreground regions. These methods [33] have been extensively studied and improved over the years. Some research in this area has focused on this topic for the specific case of thermal imagery [7, 10], while some research fuses information from infrared and the visible spectrum [9, 27]. Drawbacks of systems that rely on person detection by foreground segmentation 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 [5, 31]. Both problems can be solved by using a dedicated object detector to find people in images.

Recent advances in object detection in the visible spectrum [8, 13, 24, 32, 36, 38] encourage the application of these trainable, class-specific object detectors to thermal data. Although person detection in infrared has its own advantages as well as disadvantages when compared to detection in the visible spectrum [12], most principles can be transferred from the visible spectrum to infrared. While some techniques like the Histogram of Oriented Gradients (HOG) [8] can be directly be transferred to infrared data [34], a lot of research focuses specifically on person detection in infrared. Nanda and Davis [30] use a template based approach which builds on training samples of persons to detect person in infrared data. In [39], Xu and Fujimura use a SVM which also builds on size normalized person samples to detect and track persons. In [6], Bertozzi et al. detect pedestrians from a moving vehicle by localization of symmetrical objects with specific size and aspect ratio, combined with a set of matches filters.

For most high-level applications like situation assessment, the person detection results alone are not sufficient since they only provide a snapshot of a single point in time. For these higher level interpretation purposes, meaningful person trajectories have to be built by a tracking process. To benefit from the advantages of the dedicated object detectors, a lot of approaches directly built on the results of these person detectors to conduct tracking: Andriluka et al. introduced a method of combining tracking and detection of people in [1]. This approach uses knowledge of the walking cycle of a person to predict a persons position and control the detection. Another extension of this work [24] was proposed in [26] where a tracking was set up on the ISM based object detector. In [25] Leibe et al. further extended the work to track people from a moving camera. Gammeter et al. [14] built the tracking based on the object detector and additional depth cues obtained from a stereo camera to track people in street scenes from a moving camera. In [15], Gavrila and Munder proposed a multi cue pedestrian detection and tracking system that is applicable from a moving vehicle too. They use a cascade of detection modules that involves complementary information including stereo. Wu and Nevatia [38] introduced a system that detects body parts by a combination of edgelet features and combines the responses of the part detectors to compute the likelihood of the presence of a person. The tracking is conducted by a combination of associating detection results to trajectories and search for persons with mean shift. In both cases, an appearance model which is based on color is used for data association in tracking.

In infrared data, person tracking is a more challenging problem than in the visible spectrum. This is due to similar appearance of persons in infrared which makes identity maintenance in tracking much more difficult compared to the visible spectrum where rich texture and color is available to distinguish persons. Especially on moving cameras, where the image position of people is unstable and thus not sufficient to correctly maintain object identities, the above mentioned approaches would not be capable to track persons robustly. This is due to the different assumptions the approaches make on the availability of color, a stationary camera or special sensors like a stereo camera. An approach which focuses on pedestrian tracking without making these assumption is presented in [39] by Xu and Fujimura. Here, the tracking is built on the infrared person detection results of the SVM classifier. For that they use a Kalman filter to predict a persons position and combine this with a mean shift tracking.

In this chapter, we seize on the task of detecting and tracking multiple objects in real-world environments from a possibly moving, monocular infrared camera and by that pursue the work presented in [20, 21]. Although we focus on detecting and tracking people, our approach works independently of object specifics and is thus generically applicable for tracking any object class.

Unlike most of the before mentioned approaches we do not make any assumptions on application scenario, environment or sensor specifics. Our whole detection and tracking approach is solely built on local image features (see [35] for an extensive overview) which are perfectly suited for this task since they are available in every sensor domain. As local features, we picked SURF [2] (replaceable with SIFT [28]) features since, in our application, they have some major advantages compared to other local features like a combination of Harris keypoints [17] and shape descriptors [3] (as used in [23]).

On the keypoint level, SURF features respond to blob-like structures rather than to edges, which makes them well suited for infrared person detection since people here appear as lighter blobs on darker background (or inverted, dependent on sensor data interpretation). This is due to the use of a hessian matrix based keypoint detector (Difference of Gaussian which approximates the Laplacian of Gaussian in case of SIFT) which responds to blobs rather than to corners and edges like, e.g., Harris based keypoint detectors. The SURF descriptor is able to capture two things which are important in detection and tracking. It captures the shape of a region which is important in the training of the general person detector, because the shape of person is alike for all people. Second, it is able to capture texture (which still might be available despite infrared characteristics) properties of the regions which is important in tracking where different persons have to be distinguished from each other. Another important property is the ability of the descriptor to distinguish between light blobs on dark background and dark blobs on light background. This makes it perfectly suited for detecting people in thermal data because those here usually appear lighter than the background (or darker, dependent on sensor data interpretation).

Our detection approach is built on the Implicit Shape Model (ISM) based approach introduced in [24]. Here, a general appearance codebook is learned based on training samples. Additionally to just detecting persons as a compound, we show how this local feature based person detector can be used to classify a person's body parts, which can be input to further articulation interpretation approaches. For tracking, we introduce a novel technique that is directly integrated into the ISM based detection and needs no further assumptions on the objects to be tracked. Here, we unite object tracking and detection in a single process and thereby address the tracking problem while enhancing the detection performance. The coupling of tracking and detection is carried out by a projection of expectations resulting from tracking into the detection on the feature level. This approach is suited to automatically combine new evidence resulting from sensor data with expectations gathered in the past. By that, we address the major problems that exist in tracking: we automatically preserve object identity by integrating expectation into detection, and, by using the normal codebook-matching procedure, we automatically integrate new data evidence into existing hypotheses. The projection of expectation thus stabilizes detection itself and reduces the problem of multiple detections generated by a single real world object. By adapting the weights of projected features over time, we automatically take the history and former reliability of a hypothesis into account and therefore get by without a special approach to assess the reliability of a tracked hypothesis. Using this reliability assessment, tracks are automatically initialized and terminated in detection.

We evaluate both, the standalone person detector and the person tracking approach. The person detector is evaluated in three thermal image sequences with a total of 2,535 person occurrences. These image sequences cover the complete range of difficulties in person detection, i.e., people appearing at different scales, visible from different viewpoints, and occluding each other. The person tracking is evaluated in these three and two additional image sequences under two main aspects. First, we show how tracking increases detection performance in the first three image sequences. Second we show how our approach is able to perform tracking in difficult situations where people move beside each other and the camera is moving. Additionally, we show that the tracking is even able to track people correctly in cases where strong camera motion occurs.

This chapter is structured as follows. Section 2 covers the standalone person detection. Here, we start by introducing the detection approach in Sect. 2.1. The body part classification is described in Sect. 2.2. Section 2.3 provides an evaluation of person detection. Person tracking is discussed in Sect. 3. This section includes the introduction of our tracking approach in Sect. 3.1, the tracking evaluation in Sect. 3.2 and the tackling of strong camera motion in tracking in Sect. 3.3. Section 4 closes this chapter with a conclusion.

2 Person Detection

This section focuses on person detection. It introduces the detection technique, shows how this can be employed to classify a person's body parts and presents experimental results.

2.1 Local Feature Based Person Detection

The person detection approach we use here is based on the trainable ISM object detection approach introduced in [24]. In this section, we briefly describe the training and detection approach and the enhancements we made.

2.1.1 Training

In the training stage, a specific object class is trained on the basis of annotated sample 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.

The SURF 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 (see Fig. 1a) and is used to vote for object center locations in the detection process. This ISM is built by comparing every training feature to each

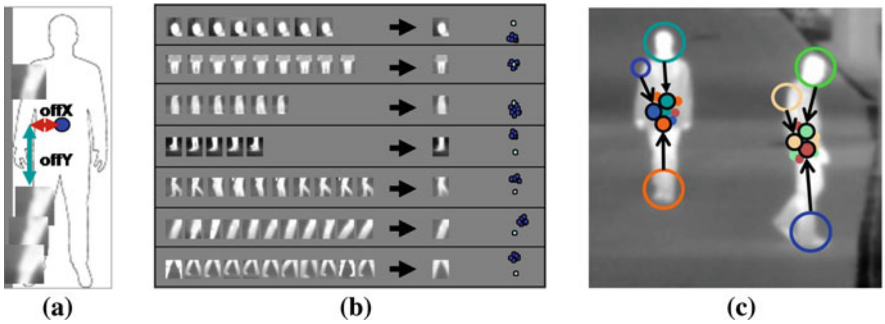

Fig. 1 **a** ISM describes spatial configuration of features relative to object center. **b** Clustered training features are mapped to a prototype. Each codebook entry contains a prototype and the spatial distribution of features. **c** Image features that match to codebook prototypes cast votes for object center locations. Each image feature has only a single vote in the final detection set since a single image feature can only provide evidence for one object hypothesis

prototype (cluster center) that was generated in the previous clustering step. If the similarity (euclidean distance in descriptor space) of a feature and the prototype 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 (Fig. 1b) entry with an assignment probability. This probability is based on descriptor similarity and a single feature can contribute to more than one codebook entry (fuzzy assignment).

2.1.2 Detection

To detect objects of the trained class in an input image, again SURF features are extracted. These features (the descriptors) are then matched with the codebook, where codebook entries with a distance below a threshold t_{sim} are activated and cast votes for object center locations (Fig. 1c). 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 [24]. First, we do not distribute vote weights equally over all features and codebook entries but use feature similarities to determine the assignment probabilities. By that, features which are more similar to codebook entries have more influence in object center voting. The assignment strength $p(C_i\,|\,f_k)$ of an image feature f_k, codebook entry C_i combination is determined by:

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

where $\rho(f_k, C_i)$ is the euclidean distance in descriptor space. Since all features with a distance above or equal t_{sim} have been rejected before, $p(C_i|f_k)$ is in range [0, 1]. The maximal assignment strength 1 is reached when the euclidean distance is 0. The same distance measure is used for the weight $p(V_{\vec{x}}|C_i)$ 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 prototype and a training feature that contributes to the codebook entry. The overall 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 approach 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 training data. Features that frequently occurred in training data generate codebook entries that comprise many offsets. A single feature (in detection) that matches with the codebook prototype thus casts many votes in object center voting with the evidence of only 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 score, 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 (see Fig. 1c). 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 [24] for details), which is necessary because objects at higher scales can be expected to generate much more features than those at 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 Sect. 2.2.

The result of the detection step is a set Φ of object hypotheses, 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.

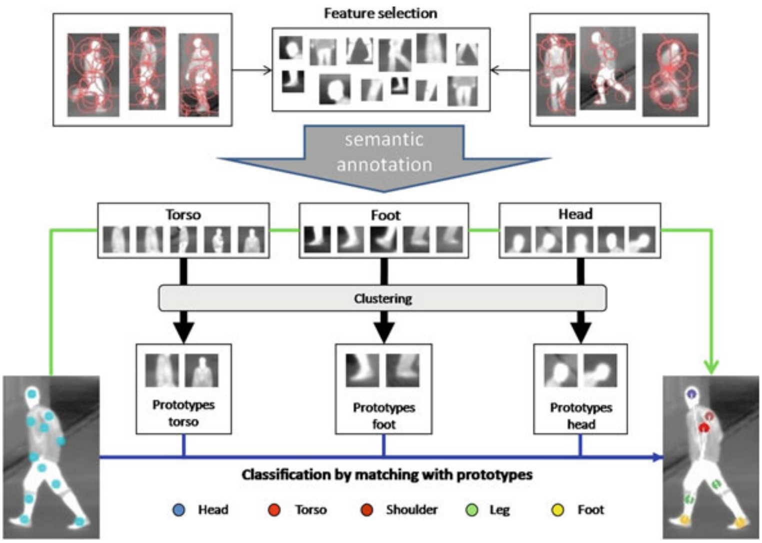

Fig. 2 Procedure of body part classification. Features found on body parts are annotated with the appropriate semantics, feature descriptors are then clustered to build appearance prototypes of each body part. Body part classification happens in two ways, the *top line* denotes the way of direct classification using the training annotation. The *bottom line* denotes classification by matching with the appearance prototypes

2.2 Body Part Classification

As mentioned in Sect. 2.1.2, 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 codebook entries for features that can be associated with certain body parts. Object hypotheses resulting from detection consist of a number of votes. The votes were generated by specific offsets (which refer to training features) in certain codebook entries which were activated by image features. As outlined in Fig. 2, using the annotation of these entries, we are now able to infer the semantics of 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 codebook representative and the image feature (see Eq. 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 training data, i.e., we pick all features annotated with "foot" from 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 to 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 descriptors 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 general this appearance is (since it was seen on many different people). 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.

Example results of the body part classification are shown in Fig. 3. Here, the relevant body part categories are: head, torso, shoulder, leg, and foot. 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 of the person's articulation. (Compare [22] for work in this direction.)

2.3 Experimental Results

2.3.1 Training Data

A crucial point in the performance of a trainable object detector is the choice of training data. 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 640 × 480. The set contains eight different persons appearing at multiple scales and viewpoints. 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

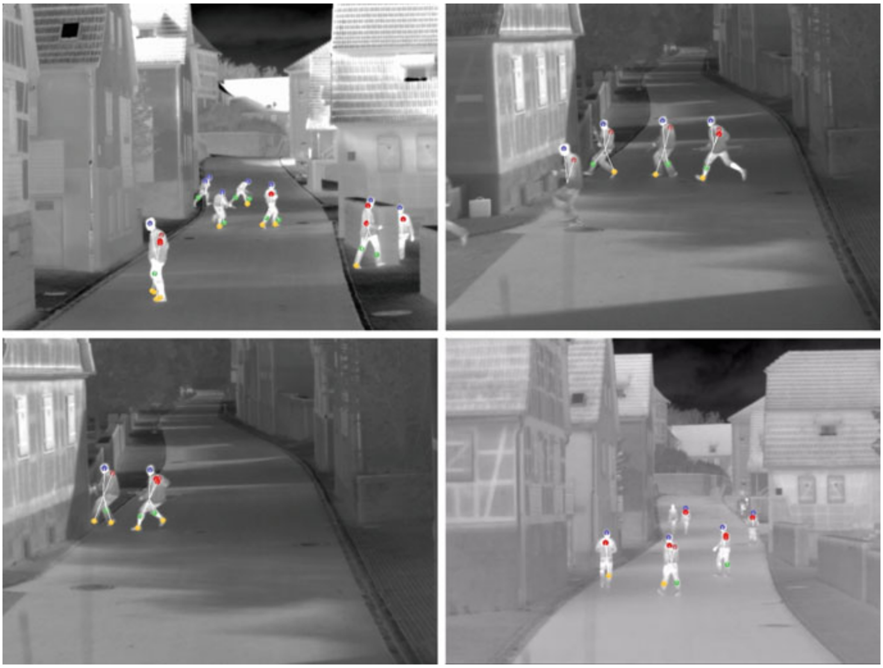

Fig. 3 Example body part classification results of detected persons. Relevant body part classes are: head, torso, shoulder, leg, and foot

(when a feature visually refers to a certain body part). Example results for the body part detection are shown in Fig. 3. All detection results shown hereafter do not contain any of the persons that appear in training data.

2.3.2 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 [25]. 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 [19] (compare intersection-over-union criterion):

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

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 which 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 Fig. 5a. 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 the top row of Fig. 4.

The second image sequence is from OTCBVS dataset [9] 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 Fig. 5b, 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 Fig. 5d 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

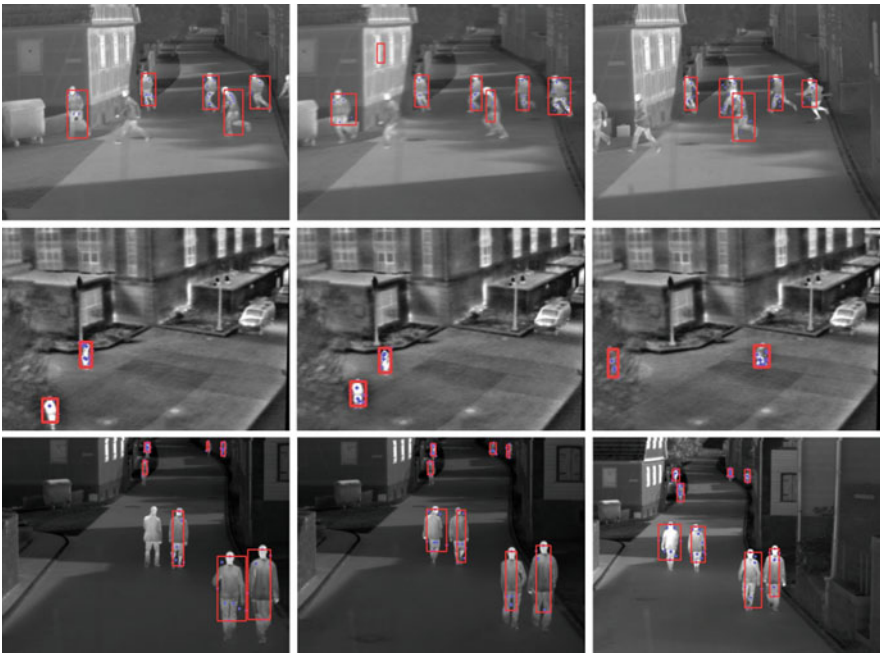

Fig. 4 Example detections of all three test sequences. Sequence 1: *top row*, sequence 2: *middle row*, sequence 3: *bottom row*. *Dots* indicate features that generate the hypothesis marked with the *bounding box*

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 the second row of Fig. 4.

The third image sequence was taken in urban terrain from a camera installed on a moving vehicle. This image sequence, with a total of 1,471 person occurrences, 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 the bottom row of Fig. 4, 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 Fig. 5c. 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—more 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 inaccuracies in bounding boxes

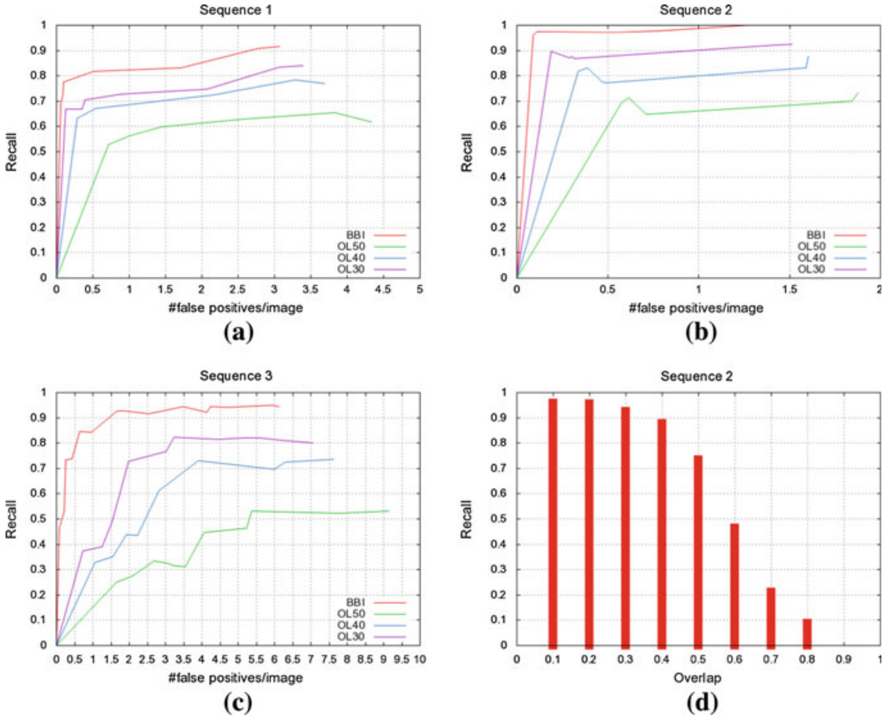

Fig. 5 Recall/false positive curves for **a** sequence 1, **b** sequence 2, and **c** sequence 3. Each chart contains four curves that refer to the different evaluation criteria. BBI: inside bounding box criterion. OL30/40/50: bounding box overlap criterion with 30, 40 and 50% overlap demand. **d** Trend 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)

of persons appearing at higher scales. This is also visible in the example detections in the bottom row of Fig. 4. Here, people in the scene background are most often detected accurately while persons close to the camera are detected rather imprecisely in terms of exact bounding boxes.

3 Person Tracking

Even a perfectly working person detector gives only a snapshot image of the surrounding. For most applications, like driver assistance or visual surveillance, it is necessary to interpret the situation over a time interval, i.e., to know where people are walking and thus know if they are a possible thread (spec. in military applications) or if we (as a driver of a vehicle) might be a thread to the person. For this, a person tracking is necessary. An important point in tracking is to consistently maintain object identities because this is a prerequisite for correct trajectory estimation. This is a difficult problem specifically in infrared data, where features like color that are commonly used to distinguish persons in tracking are not available. Here, people usually appear as a light region on a darker background which means the appearance of different persons is very alike. Additional difficulties arise when tracking should be conducted from a moving camera. In this case the use of position information for correct trajectory estimation is problematic since the camera motion distorts estimation of people motion.

In this section, we introduce a tracking strategy which is based on the object detector introduced in Sect. 2 and copes with the difficulties for tracking in infrared from a moving camera.

3.1 Local Feature Based Integration of Tracking and Detection

The object detection approach described up to now works exclusively data-driven by extracting features bottom-up from input images. At this point, we introduce a tracking technique that integrates expectations into this data-driven approach. The starting point of tracking are the results of the object detector applied to the first image of an image sequence. These initial object hypotheses build the basis for the object tracking in the sequel. Each of these hypotheses consists of a set of image features which generated the according detection. These features are employed to realize a feature based object-tracking.

3.1.1 Projection of Object Hypotheses

For every new image of the image sequence, all hypotheses Γ known in the system at this time T, each comprising a set of features Π_γ, are fed back to the object detection before executing the detection procedure. For the input image, the feature extraction is performed, resulting in a set of image features Π_{img}. For every object hypothesis in the system, the feature set Π_γ of this hypothesis γ is projected into the image. For that, we predict the feature's image positions for the current point in time (a Kalman-Filter that models the object-center dynamics assuming constant object acceleration is used to determine position prediction for features. Note that this is thought to be a weak assumption on object dynamics) and subjoin these feature to the image features.

In this joining, three different feature types are generated: The first feature type, the *native image features* Π_{img} refers to features that are directly extracted from the input image. These features contribute with the weight $P_{\text{type=nat}}$, which is set to 1.

The second feature type, the *native hypothesis features*, is generated by projecting the hypothesis features Π_γ to the image. These features are weighted with $P_{\text{type=hyp}}$ and are added to the detection-feature-set Π_γ^{tot} of hypothesis γ:

$$\Pi_\gamma^{\text{tot}} = \Pi_{\text{img}} \cup \Pi_\gamma. \tag{6}$$

These features integrate expectation into detection and their weight is set to a value in the range [0–1].

The next step generates the features of the third type, the *hypothesis features with image feature correspondence*. For this purpose, the hypothesis features Π_γ are matched (similarity is determined by an euclidean distance measure) with the image features Π_{img}. Since (i) the assignment of hypothesis to image features includes dependencies between assignments and since (ii) a single hypothesis feature can only be assigned to one image feature (and vice versa), a simple "best match" assignment is not applicable. We thus solve the feature assignment problem by the revised Hungarian method presented by Munkres in [29]. By that the best overall matching assignment and mutual exclusivity is ensured.

Feature assignments with a distance (in descriptor space) exceeding an assignment threshold κ_{feat} are prohibited. An additional image-distance constraint for feature pairs ensures the spatial consistency of features. Every $\iota \in \Pi_{\text{img}}$ which has a $\pi \in \Pi_\gamma$ assigned, is labeled as feature type 3 and contributes with the weight $P_{\text{type=mat}}$ (the matching hypothesis feature π is removed from the detection set: $\Pi_\gamma^{\text{tot}} = \Pi_\gamma^{\text{tot}} \setminus \pi$ to not count features twice). This weight is set to a value >1, because this feature type indicates conformity of expectation and data and thus contributes with the highest strength in the voting procedure.

The feature-type-weight is integrated into the voting by extending the vote weight (see Eq. 2) with factor P_{type} to

$$V_{\vec{x}}^{w} = p(C_i|f_k) \cdot p(V_{\vec{x}}|C_i) \cdot P_{\text{type}}. \tag{7}$$

The voting procedure—which is the essential point in object detection—is thus extended by integrating the three different feature types that contribute with different strengths. The whole procedure is shown in Fig. 6.

3.1.2 Coupled Tracking and Detection

From now on, the detection is executed following the general scheme described in Sect. 2. In addition to the newly integrated weight factor, the main difference to the standard detection is that the voting space contains some votes which vote exclusively for a specific object hypothesis. Besides, votes which were generated from native image features can vote for any hypothesis. This is shown in Fig. 7a. Here, different gray values visualize affiliation to different hypotheses.

Since the number and position of expected object hypotheses is known, no additional maxima search is necessary to search for known objects in the voting space. As we see in Fig. 7b, the mean shift search can be started immediately since the expected position of a hypothesis in voting space is known (the position is determined by a prediction using a Kalman filter that models object center dynamics). Starting from

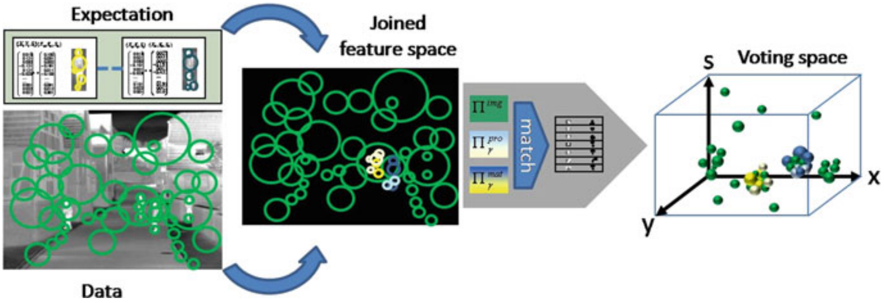

Fig. 6 Coupling of expectation and data for tracking. Features in object hypotheses (results of former detections) are propagated to the next frame and combined with new image features in a joined feature space. This feature space contains three different feature types which are generated by matching expectation and data. Π^{img}: native image features without correspondence in the hypothesis feature set, Π^{pro}: features of projected hypotheses without image feature match, Π^{mat}: matches of hypothesis and image features. The projected and matching features are marked with *grey values* according to the different hypotheses. These features can only vote for the hypotheses they refer to. The joined feature set is then input to the standard object detection approach where features are matched with the codebook to generate the voting space. Here, votes produced by native image features can vote for any object hypothesis while hypothesis specific votes are bound to a specific hypothesis

this position, the mean shift search is conducted determining the new object position. Since a mean shift search was started for every known object in particular, the search procedure knows which object it is looking for and thus only includes votes for this specific object and native votes into its search. By that hypothesis specific search, identity preservation is automatically included in the detection procedure without any additional strategy to assign detections to hypotheses. After mean shift execution, object hypotheses are updated with the newly gathered information. Since, by the propagation of the features, old and new information is already combined in the voting space, the object information in the tracking system can be replaced with the new information without any further calculations or matching.

To detect new objects, a search comprising the standard maximum search has to be conducted since the positions of new objects are not known beforehand. As we see in Fig. 7c, this maxima search is executed in a reduced voting space where only native votes that have not been assigned to a hypothesis yet remain. All votes that already contributed to an object hypothesis before are removed from the voting space. This ensures that no “double” object hypotheses are generated and determines that new image features are more likely assigned to existing object hypotheses than to new ones.

As in the original voting procedure, the initial “grid maxima” are refined with mean shift as we see in Fig. 7d. All maxima with a sufficient score found here initialize new tracks.