2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

# People Detection and Tracking from Aerial Thermal Views

Jan Portmann, Simon Lynen, Margarita Chli and Roland Siegwart
Autonomous Systems Lab, ETH Zurich

*Abstract*— **Detection and tracking of people in visible-light images has been subject to extensive research in the past decades with applications ranging from surveillance to search-and-rescue. Following the growing availability of thermal cameras and the distinctive thermal signature of humans, research effort has been focusing on developing people detection and tracking methodologies applicable to this sensing modality. However, a plethora of challenges arise on the transition from visible-light to thermal images, especially with the recent trend of employing thermal cameras onboard aerial platforms (e.g. in search-and-rescue research) capturing oblique views of the scenery. This paper presents a new, publicly available dataset of annotated thermal image sequences, posing a multitude of challenges for people detection and tracking. Moreover, we propose a new particle filter based framework for tracking people in aerial thermal images. Finally, we evaluate the performance of this pipeline on our dataset, incorporating a selection of relevant, state-of-the-art methods and present a comprehensive discussion of the merits spawning from our study.**

## I. Introduction

With the relaxation of the almost exclusive use of thermal cameras on military applications, the research community has been experiencing growing interest in their use in applications such as search-and-rescue. While people tracking in visible-light images has been studied for decades, the application of the developed methodologies on thermal images is far from straight-forward. As visual appearance cues in form of color and texture are no longer available, optaining meaningful segmentation results becomes challenging, especially given variable environmental conditions (e.g., weather, type of scenery). Beause texture information from color is rare, associating detections of humans before and after their paths cross is no longer possible unless tracking is employed on a sequence of consecutive images.

With the increasing resolution and decreasing size, weight, cost and power consumption of thermal cameras, mounting them onboard aerial platforms for search-and-rescue scenarios has become increasingly popular where victims often need to be searched for within a potentially large area. In order to boost effectiveness of rescue missions, we aim for employing Unmanned Aerial Vehicles (UAVs) to gain an overview of the scene. However, people detection from such oblique, top-down views is a real challenge and little work exists in the literature addressing this problem, especially in the context of thermal imaging.

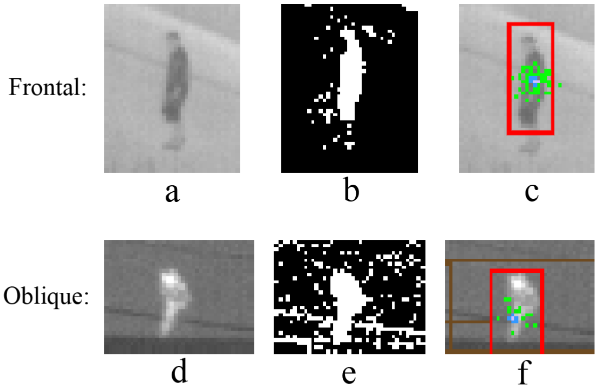

Fig. 1: Intermediate results within our framework, operating on both frontal (top row) and oblique (bottom row) views of humans. The raw image (in (a) and (d)) is background-segmented (in (b) and (e)), followed by the application of the proposed Body Part Detector (BPD) and the particle filter tracker (in (c) and (f), respectively). Black: background, white: foreground, brown: foreground blob, blue: tracker, green: particles, red: detection.

We propose a detection and tracking system processing thermal image sequences viewing the scene from viewpoints resembling those captured from of a UAV. As a novelty, we introduce a pipeline containing a robust background subtraction method and a particle filter guided detector. We show that our tracking framework outperforms all implemented detectors in terms of recall rates at high precision even if the camera setup experiences large, unsteady motion. Our framework achieves dramatic improvement on recall (of about 7× the recall of the best detector, on occasions), while reaching real-time performance of 16 Hz. Fig. 1 shows intermediate steps within our tracking pipeline when operating on both frontal and oblique views of humans present in our manually annotated and publicly available dataset.

### A. *People Detection & Tracking in visible-light images*

Detecting and tracking people, and more generally, objects, has been a very active area of research over the last couple of decades. The works in [1] and [2] present a thorough study of state-of-the-art people detectors. When an image sequence is available, tracking on top of pure detection can provide better performance as discussed in [3]. One of the first object detection approaches was proposed by Viola and Jones [4], initially for face detection. Inspired by [4] in this paper we adapt and extend this methodology to construct a body part detector in the people detection paradigm, as discussed in Section II-B. Another extremely popular detection algorithm is based on Histograms of Oriented Gradients (HOG) [5] which allows the comparison of an object shape with a pre-trained model. An extension has been proposed

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7) under grant agreements n.285417 (ICARUS) and n.600958 (SHERPA).

978-1-4799-3684-7/14/$31.00 ©2014 IEEE

in [6], which takes body parts and their relative position into account, however at the cost of lower processing speed [1].

### B. From visible-light to thermal image sequences

The application of such detectors on thermal imagery however poses several challenges: Thermal cameras still have significantly lower resolution than their visible-light counterparts and are commonly corrupted by thermal noise. Additionally, reasoning between different, correctly classified objects becomes more difficult since texture information is rare. In [7], thermal as well as visible-light images are combined to detect cars and humans from a UAV viewpoint. In [8], thermal cameras have been applied to track pedestrians from an upfront viewpoint for night driving using a fusion of the hyper permutation network, a hierarchical contour matching algorithm and a cascaded classifier, respectively. A detector based on SURF features aided by a Kalman filter to predict the motion of individual features has been proposed in [9].

Since humans most often have a different body temperature than the surrounding background, background subtraction method offers a first and fast selection method to truncate the detector search space to image regions containing humans. In [10] and [11], a background subtraction method has been applied, which requires a static thermal camera for surveillance scenarios. In this work, we employ the ViBe [12] background estimation, which is capable of segmenting regions that are both hotter and colder than the environment and can deal with situations where the camera is moving, which is essential in our UAV application. Our algorithm processes thermal (long-infrared) wavelengths only, enabling tracking during night and in other situations where the image quality of visible-light cameras is limited. As the framework is intended to be carried by a UAV, we focus on camera scenes from an elevated platform, where humans appear and move differently in image space than when observed from a ground based viewpoint. Our particle filter follows the approach of [13] with several necessary adjustments described in section Section II-C to enable tracking in thermal rather than visible imagery and on the limited computational power that can be carried by a small-sized UAV.

## II. Methodology

Our approach comprises of three steps. Background subtraction is used to generate candidate foreground regions for accelerated detection and improved tracking performance. We employ several detectors in our pipeline and evaluate their performance. Finally, the tracker uses both detections as well as foreground regions classified by the first step as guidance.

### A. Background Subtraction

Background subtraction effectively reduces the search space for the people detector, which is otherwise the most time consuming part of the tracking process, as evident in Table I. The employed background subtraction method ViBe [12] stores a background model by randomly selecting image values at the current pixel location from past frames and image values at the location of neighbouring pixels in the current image. Every pixel of a new image is classified as background if its intensity value is within a predefined threshold of the background model pixels. Connected foreground regions of a number of pixels above a threshold, form ROIs. We adapt our implementation to allow processing of 16-bit images and introduce a variable threshold which is based on the standard deviation over the image instead of a fixed threshold.

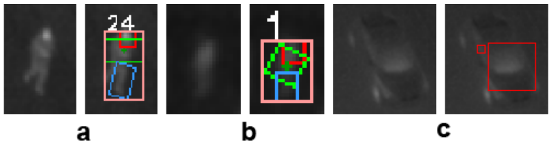

Fig. 2: Example of pairs of raw images with the obtained detections using the Body Part Detector; in (a) is a human in oblique view, in (b) the human occupies a small image region, and in (c) is a true negative detection. Color-coding of rectangles: red - head, green - upper body, blue - legs, pink - whole human.

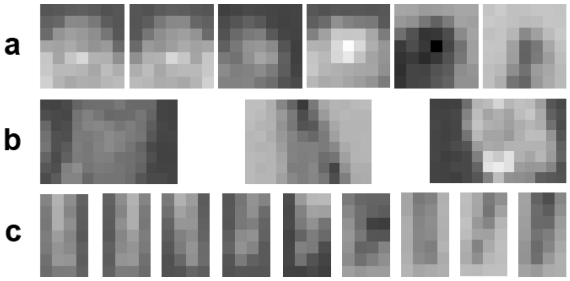

Fig. 3: Examples of training images for the different body parts in the BPD, enlarged for illustration without interpolation. In (a) are training images for heads, (b) for upper bodies, and (c) for legs.

### B. Detector

Within our tracker framework, we employ different detectors publicly available in OpenCV and propose a part based detector. For the high performance [1] HOG detector a descriptor is obtained by calculating histograms of image gradients followed by ordering and normalizing them in blocks followed by e.g. a SVM based training. A sliding window approach, is then commonly used for object detection. We trained our own HOG classifier using a linear SVM and our training dataset described in Section III with a cell size of 8×8, a block size of 16×16 px and 128×64 px sized training images. Additionally we implemented a HOG classifier trained on the well-known INRIA [5] dataset, which contains visible-light images only. An extension to HOG, LatentSVM [14] detects not only objects in the exact shape it has been trained for, but also in different body configurations in the case of articulated objects such as humans. Instead of classifying an object as a whole, LatentSVM searches for distinctive parts and returns a confidence value, taking into account the positions of the detected parts. As for HOG, we trained a detector with our own thermal dataset as well as the INRIA dataset using the parameters suggested by the authors.

Another detector concept is based on cascades of Local Binary Patterns (LBP) [15]. Here the image intensity of points around a center pixel are compared to the intensity of the center pixel itself the binary test results are accumulated in a histogram within a cell of pixels. In the training step, boosting techniques are applied to select the most representative features. Combining these tests in a cascade allow high detection rates by gradually filtering out non matching objects. Object classification of input images is obtained by using a sliding window approach and a pyramid representation of the image. Here, the LBP cascade is trained with our thermal training set.

To render cascades of simple features more competitive, we extend a Haar feature based cascade to what we coin as the “Body Part based Detector (BPD)”. Haar features are composed of rectangular windows containing two image regions, over which the pixel intensities are added; the difference of these sums correspond to the feature we use. As in [4], distinctive features are selected from positive training images using AdaBoost [16], tested against negative training images and stored in a cascade. To cope with the varying sizes and poses of humans observed from a UAV perspective, we only choose to train on head, upper body and legs as the most characteristic parts potentially observable.

After finding head candidates by applying a sliding window approach, new image scan windows are constructed by rotating rectangles around the candidate followed by applying a second detector trained for upper bodies. If the response is above a threshold, the procedure is repeated to search for a single leg. Thus, the detector becomes rotation-invariant and able to classify humans at different angles. Furthermore the employed individual detectors remain simple and can be trained very quickly, as opposed to diversifying the fundamental features. The overall detection speed outperforms other detectors such as HOG (see Table I). Fig. 3 illustrates training examples, while Fig. 2 shows example detections of our BPD.

### *C. Tracker*

A tracker increases considerably the identification and location accuracy of an object making use of both detection results and temporal constraints. Our approach is based on a particle filter, such that each identified object is assigned a number of particles. Following the assignment of detections in a new image to the corresponding particle clouds of the tracker, the particles are attributed a weight based on how well they represent the actual object. These weights are then sampled from a probability distribution used to determine the future location of the object.

Each tracker $\mathcal{T}_L$ is composed of a label $L$, $K$ particles $\mathcal{P}$, the last associated guidance candidate $\mathcal{C}$ defined later in Eq. (2), a rectangle $\mathbf{r}$ of dimensions corresponding to the average of the last $N_r$ associated guidance candidates, as well as the position $\mathbf{m}$ and velocity $\mathbf{u}$ of the tracker (mean of the particles):

$$\mathcal{T}_L : \{L, \mathcal{P} = \{P_0, P_1, ..., P_K\}, \mathbf{m}, \mathbf{u}, \mathcal{C}, \mathbf{r}\} \tag{1}$$

*1) Update:* We chose a constant velocity model to describe the propagation of the particles $\mathcal{P}$ for each tracker where the process noise for the position and velocity are drawn from a zero-mean normal distribution $\mathcal{N}(0, \frac{\sigma^2_{x,v}}{\kappa})$, where $\kappa$ is the number of successfully associated frames up to a threshold. This allows the tracker to close in to the object movement behaviour. Additionally, each particle carries a weight $w$ to indicate the likelihood of the particle correctly tracking the object. To account for rough camera movement and the associated error in velocity estimation, particle prediction is driven by a homography obtained from frame to frame optical flow tracking.

*2) Association:* Detections, as well as ROIs retrieved from the detectors in Section II-B and the background subtraction in Section II-A are merged into tracker guidance candidates $\mathcal{C}_i$ containing a position $\mathbf{m}$, a rectangle $\mathbf{r}$ and a weight obtained by the detector.

$$\mathcal{C}_i : \{\mathbf{m}, \mathbf{r}, weight\} \tag{2}$$

We divide large foreground regions containing smaller detections into two separate candidates and replace similarly sized regions with their corresponding overlaying detections. Then, a matching matrix can be constructed with matching scores $S_{iL}$ for each guidance candidate $\mathcal{C}_i$ to every current tracker. The score of a $\mathcal{C}_i$ is calculated by taking into account the probability distributions on the difference in rectangle sizes $p(\Delta r_{iL})$, the distance to the last associated guidance $p(\Delta m_{iL})$, the distances to the tracker particles $p(\Delta d_{iLk})$, a score indicating how well the candidate position is in accordance to the tracker velocity $g(\mathcal{C}_i(\mathbf{m}), \mathcal{T}_L(\mathbf{m}, \mathbf{u}))$, the result of ViBe and a weighting factor on detections $\epsilon_{det}$, according to:

$$\begin{aligned} S_{iL} = p(\Delta r_{iL}) \cdot p(\Delta m_{iL}) \cdot g(\mathcal{C}_i(\mathbf{m}), \mathcal{T}_L(\mathbf{m}, \mathbf{u})) \\ \cdot \epsilon_{det} \cdot \sum_{k \in \mathcal{T}_L(\mathcal{P})}^{\mathcal{K}} p(\Delta d_{iLk}) \ , \end{aligned} \tag{3}$$

where

$$\begin{aligned} \Delta r_{iL} &= \|\mathcal{C}_i(\mathbf{r}) - \mathcal{T}_L(\mathbf{r})\|, & p(\Delta r_{iL}) &\sim \mathcal{N}(0, \sigma_r^2) \\ \Delta m_{iL} &= \|\mathcal{C}_i(\mathbf{m}) - \mathcal{T}_L(\mathcal{C}(\mathbf{m}))\|, & p(\Delta m_{iL}) &\sim \mathcal{N}(0, \sigma_m^2) \\ \Delta d_{iLk} &= \|\mathcal{C}_i(\mathbf{m}) - \mathcal{T}_L(P(\mathbf{x})_k)\|, & p(\Delta d_{iLk}) &\sim \mathcal{N}(0, \sigma_d^2). \end{aligned} \tag{4}$$

It should be noted that $\Delta r_{iL}$ is calculated by comparing the height and width of the detection and the past associated detections and $\Delta m_{iL}$ by subtracting the distance from the position of the guidance candidate to the position of last successfully associated guidance of the tracker. The distance of the position $\mathbf{x}$ of the individual tracker particle $P_k$ to the position of the candidate is named $\Delta d_{iLk}$. The weighting factor $\epsilon_{det}$ is implemented to favor detector to background subtraction output and is defined by:

$$\epsilon_{det} = \begin{cases} \epsilon & \text{if } \mathcal{C}_i \in \text{detections} \\ 1 & \text{if } \mathcal{C}_i \in \text{foreground} \end{cases} \tag{5}$$

As with all the thresholds we use in this work, the value of $\epsilon$ has been determined using nonlinear optimization on a seperate training dataset (”Sempach-11”), as described in Section III-B.

Inspired by [13], $g(\mathcal{C}_i(\mathbf{m}), \mathcal{T}_L(\mathbf{m}, \mathbf{u}))$ takes into account the velocity of the tracker according to:

$$g(\mathcal{C}_i(\mathbf{m}), \mathcal{T}_L(\mathbf{m}, \mathbf{u})) = \begin{cases} p(\Delta s_{iL})_{dist} & \text{if } \|\mathcal{T}_L(\mathbf{u})\| < \tau \\ q(\mathcal{C}_i(\mathbf{m}), \mathcal{T}_L(\mathbf{m}, \mathbf{u})) & \text{otherwise,} \end{cases} \tag{6}$$

where

$$\Delta s_{iL} = \|\mathcal{C}_i(\mathbf{m}) - \mathcal{T}_L(\mathbf{m})\|, p(\Delta s_{iL}) \sim \mathcal{N}(0, \sigma_s^2) \ ,$$

such that $\Delta s_{iL}$ represents the distance between the guidance candidate and the position of the tracker and $\tau$ is a threshold on the tracker speed. Function $q \sim \mathcal{N}(0, \sigma_q^2)$ assigns a weight based on a zero-mean Gaussian distribution by calculating the distance of the candidate location $\mathcal{C}_i(\mathbf{m})$ to the line defined by the tracker location $\mathcal{T}_L(\mathbf{m})$ and the velocity vector $\mathcal{T}_L(\mathbf{u})$, divided by $\Delta m_{iL}$ (Eq. (4)) and weighted by the norm of the difference of the normalized association vector to the candidate location and the normalized velocity vector. Only trackers that have a speed more than $\tau$ benefit from this function to suppress noise.

Having obtained the matching score matrix, the particle filter associates highest matching guidance candidates to the corresponding trackers, forming a new $\mathcal{T}_L(\mathcal{C})$.

*3) Particle Weighting:* The next step in particle filtering involves the calculation of the individual particle weights $w_{P_k}$ for each tracker $\mathcal{T}_L$:

$$w_{P_k} = p(\|(P_k(\mathbf{x})) - \mathcal{C}(\mathbf{m})\|) \cdot \varphi(P_k(\mathbf{x})) + \rho, \tag{7}$$

with $\rho \sim \mathcal{N}(0, \sigma_\rho^2)$ to prevent particle starvation due to similar $w$, the distance of the particle to the guidance $\mathcal{C}$ (zero if none associated) taking values in $p(\|(P_k(\mathbf{x})) - \mathcal{C}(\mathbf{m}))\|) \sim \mathcal{N}(0, \sigma_w^2)$ and the weighting factor $\varphi$ defined as:

$$\varphi(P_k(\mathbf{x})) = \begin{cases} \theta > 1 & \text{if } \mathbf{x}_i \in \text{foreground} \\ 1 & \text{otherwise,} \end{cases} \tag{8}$$

to increase the weight if the particle is located on foreground.

*4) Resampling:* Finally, a random number drawn from the uniform distribution from zero to the sum of all the particle weights is taken. New particles are then selected by summing up their associated weights until their sum is higher or equal to the random number.

*5) Initialization:* A new cloud of normally distributed particles around the detection center is initialized for a new tracker if an unassociated candidate detection is above a threshold, fully inside the frame and positively classified by the detector in two consecutive frames.

*6) Deletion:* To avoid ghost trackers with no associated detections or ROIs, each tracker is examined at every iteration to reside within the frame and have an association with a ROI not further further back than a fixed number of frames $n_{ROI} < n_{detection}$ and a fixed number of iterations $n_{detection}$. Association vectors $\nu_L$ between tracker $\mathcal{T}_L$ and the corresponding detection or ROI have to be steady in both direction and length, since heavily fluctuating associations are most likely caused by misguided trackers. This is decided by setting a threshold $\gamma$, which needs to be larger than the averaged dot product $\psi(\nu_L)$ of the association vectors divided by the average vector length defined as $bias(\nu_L)$. Namely, $\gamma$ is defined as:

$$\gamma > \frac{\psi(\nu_L)}{bias(\nu_L)} \tag{9}$$

$$\psi(\nu_L) = \frac{1}{2(N-1)^2} \sum_{j \in \mathcal{V}_L}^{N} \sum_{k \in \mathcal{V}_L}^{N} \begin{cases} \nu_{L,j} \cdot \nu_{L,k} & \text{if } j < k \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $\mathcal{V}_L = \{\nu_L^t, \nu_L^{t-1}, ..., \nu_L^{t-N}\}$ are the past $N$ association vectors of tracker $\mathcal{T}_L$, and $bias(\nu_L) = \frac{\| \sum_{j \in \mathcal{V}_L}^{N} \nu_{L,j} \|}{N}$.

If one of these rules applies, the corresponding tracker will be deleted. Since texture information in thermal wavelength images is sparse and therefore inter-object distinction is challenging even for the human eye, we do not store information about deleted trackers for future re-initialization.

## III. Experiments

### A. Dataset

To the best of our knowledge, the only publicly available high quality dataset containing thermal images is OTCBVS [10]. However, the images are highly non-uniformly sampled in time and thus, not suited for tracking applications.

We introduce a new dataset[1] including 4381 manually annotated images containing humans and animals (e.g., cat, horse), as well as background scenery. The dataset is composed of 9 outdoor sequences captured at a uniform sampling rate (20 Hz) from different viewing angles and at varying temperatures. For the recordings, a FLIR Tau 320 thermal camera was used (visible in Fig. 4(f)) with a 324×256 resolution, which was handheld on an elevated platform (roughly between 10-30m above ground) to replicate the top-down viewpoints from a flying UAV.

### B. Evaluation Methodology

For our evaluation, we use the three sequences ‘ETHZ-CLA”, “Sempach-7”’ and “Sempach-10” with a total of 1282 frames.

In order to train the detectors, we used a training set containing 5578 true instances of humans with different body and outdoor temperatures, background and poses taken

[1] Our publicly available dataset can be accessed on `http://projects.asl.ethz.ch/datasets/doku.php?id=ir:iricra2014`.

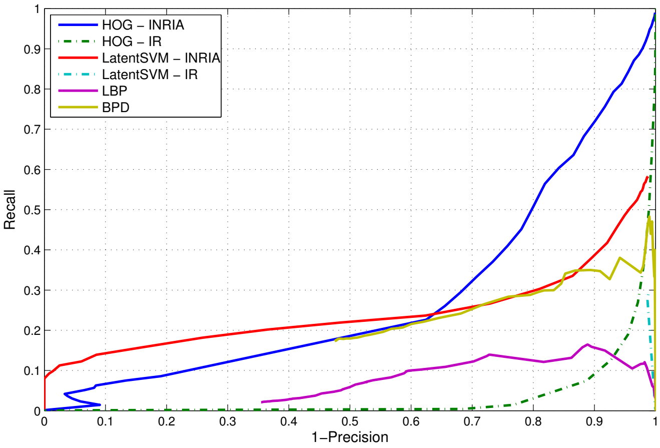

Fig. 5: Sequence: "Sempach-7". Comparison INRIA and our thermal training sets. BPD, INRIA trained HOG and LatentSVM perform best (best viewed in color).

from the "Sempach-8", "Sempach-12" and other images we recorded at different weather conditions. The whole training set is provided online.

With $A_{anno}$ denoting the area of the annotation (ground-truth) rectangle surrounding a human in the scene, and $A_{det}$ the area of the detection, a detection is classified as true positive when the following holds:

$$\frac{A_{anno} \cap A_{det}}{min\{A_{anno}, A_{det}\}} \geq 0.5. \tag{11}$$

Likewise, an annotation is regarded as successfully tracked if the position of the tracker $\mathbf{x}_{mean}$, defined by the mean of all the tracker particles $\mathcal{T}_L(\mathbf{m})$ satisfies $\mathbf{x}_{mean} \in A_{anno}$.

The various particle filter parameters $\mathbf{h}$ described in Section II-C are trained on the "Sempach-11" sequence. To achieve high recall rates at reasonable precisions, the desirable performance metric of rescue personal for people search scenarios, we formulate the optimization criteria, such that $\mathbf{h}_{max} = max\{recall(\mathbf{h}) + 0.2 \cdot precision(\mathbf{h})\}$. Finally, we produce Recall vs. 1-Precision plots by averaging the results over five passes of a sequence to limit the influence of random particle initialization and propagation addition in the particle filter.

## IV. Results

To decouple the detector evaluation from the performance of the tracker framework, we first look at the performance of the detectors trained on different training sets. Fig. 5 illustrates the relative performance of the different detectors on the "Sempach-7" sequence. Both HOG [5] and LatentSVM [6] perform significantly better when trained on the INRIA dataset [5] rather than on our thermal training set. This can be explained by the fact that the INRIA dataset only contains pictures of upright, standing people and with the typical people's height exceeding 100 px, it allows very clear training examples on the human figure. In contrast, our training set is far more complex due to the viewpoints resembling those of a UAV, containing images of people from highly oblique angles with people-heights as low as 12 px.

| Detector | Time per Frame (sec) |
|---|---|
| HOG | 0.1054 |
| BPD | 0.0384 |
| LatentSVM | 9.6057 |
| LBP | 0.074 |
| Our Particle Filter Framework | Time per Frame (sec) |
| ViBe | 0.0091 |
| Particle Filter | 0.0126 |
| Rest (resizing, blobbing, etc.) | 0.0023 |

TABLE I: Comparison of detection time per frame integrated in our particle filter on an Intel i5 @ 3.3 GHz without substantial code optimization on sequence "ETHZ-CLA". The whole tracker needs the particle filter framework and a choice of any detector.

The training stage of the detectors requires images of the same size, which necessitates up- and down-sampling for examples from our dataset. The segmentation of humans from background therefore becomes less shared and hence the magnitude of local gradients is reduced, deteriorating the discriminanility of the SVM decision boundary. Similar effects are caused by different viewpoint angles, where boundaries between human and background occur at variable pixel locations depending on which side of the camera they appear. However, the authors of [17] suggest that training with low resolution thermal images can still result in a robust classifier, which leads to the conclusion that the viewpoint angle differences are more substantial to the performance decrease.

As a result, HOG and LatentSVM are trained best on samples as they appear in the INRIA dataset, while experiencing significant decline in performance when humans are not seen from a frontal view. Even from a UAV's perspective, humans can often be captured in a nearly frontal view, e.g. when observed far away from a front-looking camera or if the ground surface is significantly inclined (e.g. in alpine rescue scenarios). Hence, detectors trained on frontal views remain a valid choice.

Since the LatentSVM detector is built from HOG detections of different body parts and therefore the resulting windows smaller than the whole body, the result is an even stronger blur on the boundaries between body and background and thus, worse performance (See Fig. 5). As in the test sequence humans are observed from various distances, the detectors experience degradation in performance with larger distance of humans from the camera. If the humans are close to the camera, the LatentSVM outperforms HOG, because the necessity of appearance similarity with the training images is reduced. However, with smaller sizes, the human form as a whole is simpler to detect than the individual body parts, which explains HOG overtaking LatentSVM in terms of recall at lower precisions, hence exposing the problem of interpolating smaller images on the detection performance of LatentSVM.

Fig. 5 shows that the standard LBP cascade [15] cannot compete with the INRIA-trained HOG, LatentSVM or our BPD, reaching at most half the recall rates at any given precision. Consequently, we focus on the integration of the better performing INRIA-trained detectors in our particle filter framework, dropping LBP cascade from further analysis.

Since processing time analysis in Table I shows that

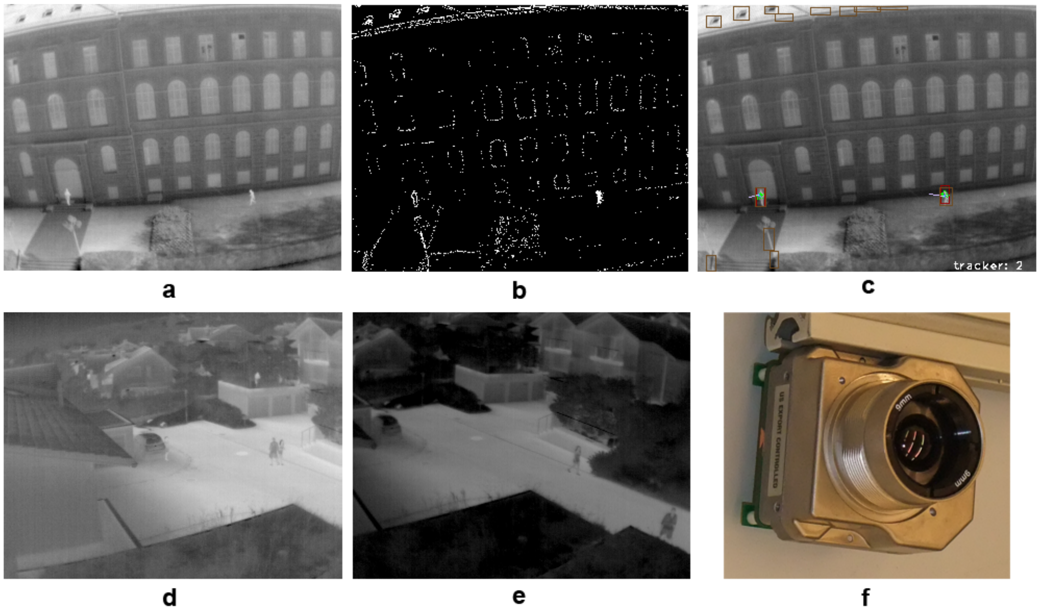

Fig. 4: The raw input image from “ETHZ-CLA” shown in (a) is fed through ViBe background segmentation to obtain (b), while (c) depicts the results of the tracker (in brown: ROI rectangles, red: detections, green: particles of the tracker, white: velocity vector). In (d) is a typical image taken from “Sempach-7”, while (e) is taken from “Sempach-10”. Finally, in (f) is the TAU FLIR 320 thermal camera used in our setup. The body temperature of the people w.r.t. the environment is higher in (a), while being lower in (d) and (e) with an exception of a single person in (d) in the top right corner.

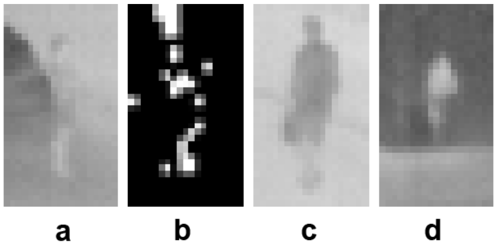

Fig. 6: From left to right: (a) human barely visible in raw image, (b) ViBe background segmentation on (a), (c) raw image of human with lower body temperature, and (c) raw image of human with higher body temperature. All snippets are taken from a single frame inside the “Sempach-7” sequence.

LatentSVM is far from able to process video frames in real-time, our further analysis focuses on the two remaining detectors HOG and BPD. The full processing pipeline, e.g. combining the BPD with our tracker, runs in real-time requiring about 65 ms per image.

The “Sempach-7” sequence features humans both far away and close to the camera (See Fig. 6). At higher precisions, the HOG-Tracker shows far superior performance than the other methods, roughly doubling the recall of the best detector (See Fig. 7). Furthermore, the BPD-Tracker can increase precision rates by a factor of two compared to the standalone BPD. The background subtraction implemented in the tracking framework filters out many potential false positives, thus increasing precision. The sequence “Sempach-7” highlights a set of challenges for our framework: with low temperature difference between humans and their environment, the background subtraction algorithm does not segment the humans, preventing the detector to label the region as a human and thereby lowering the recall value. This can be observed at lower precisions, where the detectors eventually classify barely visible humans such as the one shown in Fig. 6(a) as positive, while the HOG-Tracker does not.

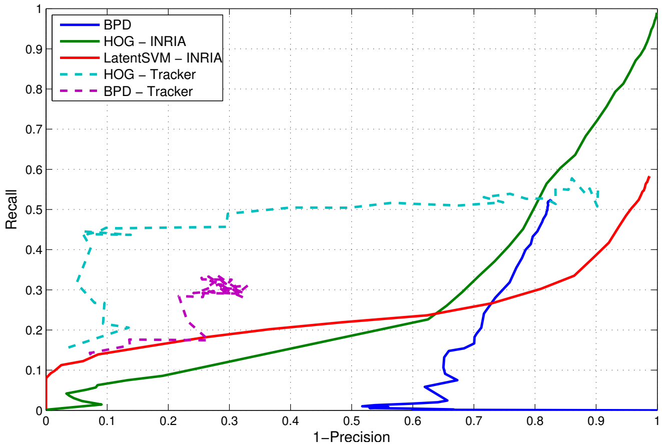

Fig. 7: Sequence: “Sempach-7”. Comparison of detection and tracking methods. Our HOG-Tracker outperforms the detectors at high precisions by a factor of two. Meandering can appear since the pipeline contains stochastic elements.

In sequence “ETHZ-CLA”, people are clearly distinguishable from the background (a typical image shown in Fig. 4(a)), therefore the tracker can exploit the full potential of the background subtraction method (Fig. 4(b)) as a guidance for the particles. As expected and evident from Fig. 8, the HOG-Tracker significantly outperforms the standalone detector in terms of recall by at least a factor of seven at higher precisions. The BPD-Tracker does not reach the performance of the HOG-Tracker, but still obtains recall rates at least twice as good as the best detector. None of the detectors reach the performance they achieve on “Sempach-7” illustrated in Fig. 7 since people in this sequence are further away from the camera, hence the body shapes become less detailed in the images, which has the greatest impact on the LatentSVM detector.

In sequence “Sempach-10” pictured in Fig. 9, constant

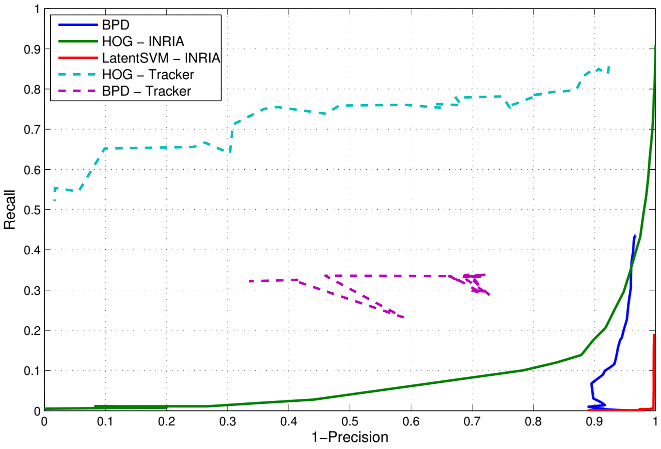

Fig. 8: Sequence: “ETHZ-CLA”. Comparison of detection and tracking methods. Both trackers heavily outperform the detectors by a factor of three respectively seven, exploiting the background subtraction algorithm.

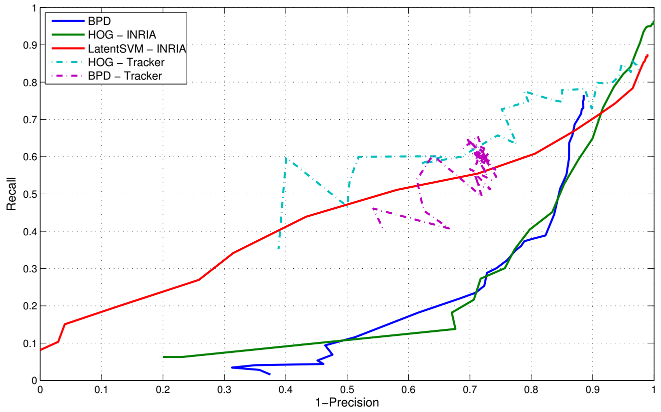

Fig. 9: Sequence: “Sempach-10”. Comparison of detection and tracking methods. Both our trackers performs similar to the much slower LatentSVM despite the heavy camera movement.

disappearance and reappearance of humans due to the heavy swifts of the camera require regular tracker initialization and deletion and the sequence is anyway challenging in terms of particle location stabilization, thus lowering the tracker performance. Here, the humans appear larger and LatentSVM can profit from detecting single body parts, thereby obtaining better recall values than the other detectors. Otherwise, HOG and BPD appear to perform roughly similar.

## V. Conclusions

In this paper, we propose a tracking algorithm based on a particle filter combined with background subtraction for people tracking in thermal-infrared images. We show that the tracker is capable of locating people in different scenarios from viewpoints such as the ones experienced by aerial platforms used in search and rescue. Using background segmentation to both reduce the detection space and to serve as guidance, the tracker substantially increases people locating accuracy and detection speed. We present a comprehensive study of existing algorithms as they are applied on our dataset. This dataset, containing thermal image sequences of urban scenery observing humans and animals from oblique, top-down viewpoints complete with annotations, is made publicly available.

Future work will focus on employing a higher resolution thermal camera to aid people detection, which remains the bottleneck in tracking algorithms. In situations where the contrast between people and background is low, true detections can be disregarded by the background subtraction. Consequently a combination with additional sensing modalities (e.g. visible-light cameras) would be necessary to disambiguate in these cases.

## VI. Acknowledgments

The authors are grateful to the help of Michael Burri, Alexandra and Bruno Portmann and Carmen Schwizer for capturing datasets.

## References

[1] P. Dollár, C. Wojek, B. Schiele, and P. Perona, “Pedestrian detection: An evaluation of the state of the art,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2011.

[2] M. Enzweiler and D. Gavrila, “Monocular pedestrian detection: Survey and experiments,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2009.

[3] M. Andriluka, S. Roth, and B. Schiele, “People-tracking-by-detection and people-detection-by-tracking,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[4] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[5] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *International Conference on Computer Vision & Pattern Recognition*, 2005.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2010.

[7] A. Gaszczak, T. P. Breckon, and J. Han, “Real-time people and vehicle detection from UAV imagery,” in *Proceedings of the SPIE Conference Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, 2011.

[8] M. Mählisch, M. Oberlander, O. Lohlein, D. Gavrila, and W. Ritter, “A multiple detector approach to low-resolution fir pedestrian recognition,” in *Proceedings of the Intelligent Vehicles Symposium (IV)*, 2005.

[9] K. Jüngling and M. Arens, “Local feature based person detection and tracking beyond the visible spectrum,” in *Machine Vision Beyond Visible Spectrum*. Springer Berlin Heidelberg, 2011, vol. 1, pp. 3–32, ISBN: 978-3-642-11567-7.

[10] J. W. Davis and M. A. Keck, “A two-stage template approach to person detection in thermal imagery,” in *IEEE Workshop on Applications of Computer Vision*, 2005.

[11] W. Wang, J. Zhang, and C. Shen, “Improved human detection and classification in thermal images,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2010.

[12] O. Barnich and M. Van Droogenbroeck, “ViBe: a powerful random technique to estimate the background in video sequences,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[13] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, “Online multiperson tracking-by-detection from a single, uncalibrated camera,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2011.

[14] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, “Discriminatively trained deformable part models, release 4,” http://people.cs.uchicago.edu/~pff/latent-release4/, implemented as LatentSVM.

[15] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Li, “Learning multi-scale block local binary patterns for face recognition,” in *Proceedings of the International Conference on Advances in Biometrics (ICB)*, 2007.

[16] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” in *Journal of Computer and System Sciences*, 1997.

[17] F. Suard, A. Rakotomamonjy, A. Bensrhair, and A. Broggi, “Pedestrian detection using infrared images and histograms of oriented gradients,” in *Proceedings of the Intelligent Vehicles Symposium (IV)*, 2006.