The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Vision Based Victim Detection from Unmanned Aerial Vehicles

Mykhaylo Andriluka[1], Paul Schnitzspan[1], Johannes Meyer[2], Stefan Kohlbrecher[1],
Karen Petersen[1], Oskar von Stryk[1], Stefan Roth[1], and Bernt Schiele[1,3]
[1]Department of Computer Science, TU Darmstadt
[2] Department of Mechanical Engineering, TU Darmstadt
[3] MPI Informatics, Saarbrücken

***Abstract*— Finding injured humans is one of the primary goals of any search and rescue operation. The aim of this paper is to address the task of automatically finding people lying on the ground in images taken from the on-board camera of an unmanned aerial vehicle (UAV).**

In this paper we evaluate various state-of-the-art visual people detection methods in the context of vision based victim detection from an UAV. The top performing approaches in this comparison are those that rely on flexible part-based representations and discriminatively trained part detectors. We discuss their strengths and weaknesses and demonstrate that by combining multiple models we can increase the reliability of the system. We also demonstrate that the detection performance can be substantially improved by integrating the height and pitch information provided by on-board sensors. Jointly these improvements allow us to significantly boost the detection performance over the current de-facto standard, which provides a substantial step towards making autonomous victim detection for UAVs practical.

I. Introduction and Related Work

Finding human victims in post-disaster scenarios is one of the primary goals of any search and rescue (SAR) operation. Although significant progress has been made in developing ground robots for SAR applications, most of these robots still lack the mobility necessary for autonomous exploration of disaster sites. However, with the emergence of lightweight and inexpensive unmanned aerial vehicles (UAVs) it becomes possible to quickly survey a disaster site from the air in order to identify humans needing help [8], [21], [29].

In this paper, we focus on victim detection from a UAV using an on-board daylight camera as our main sensing device. We envision that the development of powerful vision-based victim detection methods will lead to a reduction in the number (and weight) of required on-board sensors and result in cheaper, smaller, and more power efficient UAVs. Additionally, robust vision-based detectors have proven to be an important building block when designing human detection systems based on multiple sensor modalities [20], [32].

Detection of people in images is a challenging problem. While significant progress has been made in specialized areas such as pedestrian detection [13], most approaches work best when people are fully visible and appear in a limited range of poses such as standing or walking. The best performing methods often use a monolithic representation of people, such as a HOG descriptor [9], and discriminative classifiers. Models of this type have recently been extended to incorporate motion [10], [36] and color [35] features. They have also been applied to upper body detection [18], and have been integrated within larger systems to enable obstacle detection in mobile environments [14].

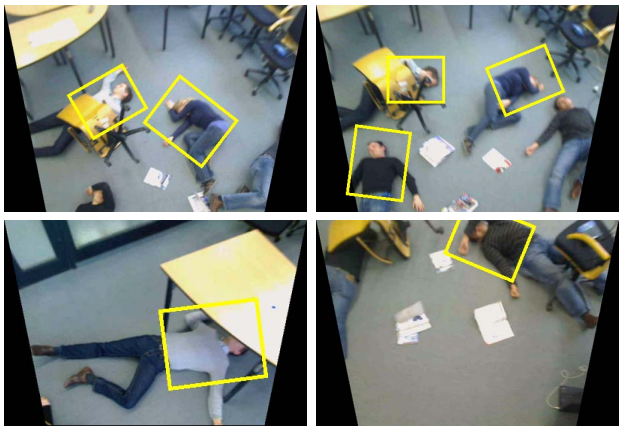

Fig. 1. Several examples of people detection obtained with our approach on images captured from a quadrotor UAV.

However, models based on monolithic representations are unlikely to generalize well to complex scenarios encountered in search and rescue applications [28]. In particular they are severely challenged by partial occlusions and high variabilities in poses of people, which frequently occur in such data. Fig. 1 shows several sample images acquired from the on-board camera of our UAV[1], which demonstrate the complexity of people detection in the scenario considered in this paper: People are partially occluded and occur in a wide range of poses in an environment, which is typically highly cluttered. Using a monolithic person representation to detect characteristic full-body shapes is prone to fail in such a scenario.

A second type of people detection methods – called part-based models in the following – proceeds by decomposing complex appearances of humans into multiple components or parts [1], [5], [17]. In [1] strong discriminatively trained body part detectors are combined with a flexible body model based on the *pictorial structures* framework, allowing to detect highly articulated people. The approach of [17] also builds on the pictorial structures framework, but introduces a training procedure based on an unsupervised discovery of model parts, which are automatically chosen to optimize

[1]Shown detections correspond to the confidence level with equal precision and recall.

detection performance. [5] proposes to address the detection of articulated and partially occluded humans using a large number of specialized part detectors that are trained to detect body regions with characteristic appearances, which are also informative for the underlying 3D body configuration. In contrast to the monolithic representations mentioned above, these part-based models seem better suited to address the challenging problem of victim detection from a UAV.

The paper makes the following contributions. First, the paper evaluates and discusses several state-of-the-art visual people detection methods on a newly recorded dataset of images taken from a UAV. This evaluation includes both methods based on monolithic and on part-based representations. In particular we consider 2 people detectors [9], [18] based on monolithic representations and 3 part-based detectors [1], [5], [17] and evaluate their suitability for detecting articulated and partially occluded people using a dataset of images seen from an on-board camera. Second, after having identified the most suitable approaches to people detection from a UAV, we propose to augment these detectors with a prior distribution based on the pitch and height of the UAV measured by the on-board sensors. We demonstrate that this prior significantly improves the performance of people detectors and analyze reasons why not all detectors equally benefit from this. The third contribution of the paper is that we demonstrate that the considered people detectors are complementary, and that by combining them we can further improve the detection performance.

The rest of this paper is organized as follows. After reviewing related work, section Sec. II describes our quadrotor UAV used for data acquisition in our experiments. In Sec. III we introduce several state-of-the art approaches to people detection, discuss their strengths and weaknesses, and describe our extensions. We present an experimental evaluation of these people detectors and our extensions in Sec. IV, and conclude and discuss future work in Sec. V.

Related Work. In addition to work on vision-based people detection discussed above, there exists a large body of literature on people detection using other types of sensors. Since many mobile robotic systems are equipped with laser range scanners, significant effort have been dedicated to using them for people detection and tracking [3], [2], [7], [19], [31]. The complementarity between visual and laser based detectors has been explored in [20], where a laser range scanner is used both to extract regions of interest in camera images and to improve the confidence of an AdaBoost based visual detector. Thermal images have also been extensively used for people detection using either specifically designed methods [11], [30] or by directly applying methods originally designed for detection of people in daylight images [33].

Leveraging complementary information of different types of sensors was recently proposed in the context of autonomous victim detection [12], [24]. The work of [12] comes particularly close to ours in that it also addresses victim detection from UAVs. The authors propose to utilize a thermal camera to pre-filter promising image locations and subsequently verify them using a visual object detector. While in [12] people lying on the ground are assumed to be in upright poses, in our paper we address the significantly more complex problem of detecting arbitrarily articulated people. Note that the results of our work can still be used in combination with thermal camera images, which similarly to [12] can be used to restrict the search to image locations likely to contain people or to prune false positives, which contain no thermal evidence.

Fig. 2. Quadrotor platform used for the experiments.

While combining multiple sensors for people detection is clearly beneficial in many scenarios it comes at the cost – in particular for unmanned aerial vehicles – of an increased payload for the additional sensors. This paper therefore aims to evaluate and push the state-of-the-art in visual people detection in order to minimize sensor requirements for this task.

II. System Overview

The platform used for our experiments is a quadrotor helicopter developed at TU Darmstadt (Fig. 2). These kinds of vehicles are able to take off and land vertically and can hover at a fixed position, which motivates their application to search and rescue missions [23]. The propulsion system using four independently controlled motors and propellers allows the carriage of comparatively heavy payloads. Our quadrotor can carry up to 500g of cameras and other sensors and weighs 1200g including the controller system and batteries for an endurance of approximately 20 minutes. With a diameter of 80cm it can be easily deployed in outdoor missions as well as for indoor scenarios.

Due to the instability of a quadrotor, the vehicle's attitude and velocity has to be controlled permanently. Therefore it is equipped with a 3-axis inertial sensor and magnetometer, a pressure sensor, a GPS receiver, and an ultrasonic ranger to measure the distance to the ground. The sensor information is fused using an extended Kalman filter running at 200 Hz deriving an integrated navigation solution using the algorithm of [34]. The outputs of the filter are fed to a cascaded PID controller to stabilize the attitude, height, velocity, and position of the quadrotor. As the rotational direction of two adjacent drives differ, the moments about all three axes and the total thrust can be controlled independently by simply varying the speed of the individual motors.

The control system is divided into two subparts: a microcontroller board interfacing the analog and digital sensors and motors; and a commercial embedded PC platform based on a current Intel Atom processor [27]. The onboard computer executes the navigation, flight control, high-level mission control and communication tasks using the OROCOS Real-Time Toolkit [6]. It interfaces the sensor board using a real-time enabled ethernet link.

For image acquisition a Logitech QuickCam Pro9000 camera is mounted to the quadrotor, which can transmit video images to the ground stations using the wireless network. Additionally, up to five frames per second are stored on an onboard flash media for after-mission analysis, including references to the available navigational data. The intrinsic camera parameters are calibrated using a publicly available calibration toolkit [4] and the extrinsic parameters relative to the ground plane are estimated using the height and attitude estimates provided by the UAV integrated navigation solution.

III. Vision-based People Detection

Detection of people in images is a challenging problem and many approaches to people detection have been proposed over the years. Approaches are often designed with a specific subproblem in mind, such as detection of pedestrians in street scenes [9], upper body detection [18], simultaneous detection and pose estimation [1], or generic people detection [5], [15].

One of the main contributions of this paper is therefore to evaluate the applicability of these methods to our scenario and subsequently focus on improving performance of the best performing methods. While the evaluation is done in the context of victim detection from a UAV we believe that its results are applicable to people detection from mobile robots in general. In addition, we also demonstrate that in the case of a UAV we can further improve detection performance by using on-board sensor measurements in order to impose a prior on the scale of people in the image. In this section we briefly describe each of the considered approaches, and present an experimental comparison in Sec. IV.

Monolithic models. One of the most popular and effective models for people detection proposed to date is the histograms of oriented gradients (HOG) detector [9]. In this model, histograms of image gradients are calculated and normalized in a local and overlapping block scheme and concatenated to a single descriptor of a detection window, which is densely scanned over all scales and locations in a test image. We consider HOG a monolithic model because the evidence of one detection window is encoded in a single descriptor, which is cast to a discriminative classifier (e.g., SVM), making a decision about presence or absence of the object of interest. This pairing with a powerful, discriminative classifier enables high levels of performance for object detection in cluttered scenes, e.g., pedestrian detection in street scenes [13]. HOG was shown to learn a robust outer shape, which is shared by the positive training instances and delimits positive from negative samples. The local, overlapping normalization scheme enables robustness to illumination changes and to small variations in viewpoint. However, in the presence of high variability in articulation and partial occlusion HOG often fails because the model cannot recover from distorted monolithic descriptors. In our evaluation we consider two variants of HOG-based methods. i) The first variant is trained on full bodies of pedestrians. We make use of the implementation of our colleagues[2] [36]. ii) The second variant is trained on upper bodies of people[3] [18]. Such monolithic approaches are a de-facto standard when it comes to detection of people in settings with relatively little pose variation. However, it remains unclear, how these models generalize to the more challenging search and rescue scenario.

Part-based models. Part-based detection gives the flexibility necessary to deal with highly varying body poses. We consider three recently proposed part-based people detection methods: discriminatively trained part based models (DPM)[4] [17], pictorial structures with discriminant part detectors (PS)[5] [1], and poselet based detection (PBD)[6] [5].

The PS detector is built on the *pictorial structures* framework introduced in [16]. Here an object is represented as a flexible configuration of parts where one such configuration is denoted by $L = \{\mathbf{l}_0, \ldots, \mathbf{l}_N\}$, with $\mathbf{l}_i$ denoting the location of part i. In this generative formulation, the posterior over part configurations L given image evidence E is obtained via Bayes' rule: $p(L|E) \propto p(L)p(E|L)$. In order to enable efficient inference, PS employs a tree-structured Gaussian prior on L, and assumes that the overall likelihood can be decomposed into the product of individual part likelihoods. Under these assumptions the configuration posterior factorizes as:

$$p(L|E) \propto p(\mathbf{l}_0) \cdot \prod_{i=0}^{N} p(E|\mathbf{l}_i) \cdot \prod_{(i,j)\in G} p(\mathbf{l}_i|\mathbf{l}_j). \qquad (1)$$

Sum-product belief propagation is applied in order to compute the marginal posterior of the torso, $p(\mathbf{l}_0|E)$, which is then used to delimit the detection bounding box. The detection results often account for multiple overlapping boxes that are post-processed with *non-maximum suppression* keeping only the hypothesis with the highest probability from significantly overlapping hypotheses.

The PS model employs body parts corresponding to upper and lower arms and legs, torso and head, and requires examples with labeled parts for training. Part likelihood terms $p(E|\mathbf{l}_i)$ are represented with discriminative part classifiers trained with AdaBoost. The pairwise terms $p(\mathbf{l}_i|\mathbf{l}_j)$ are estimated with maximum likelihood using the provided part labels.

[2]Project page: www.mis.tu-darmstadt.de/tud-brussels/

[3]Source code available at: www.robots.ox.ac.uk/~vgg/software/UpperBody/

[4]Source code available at: people.cs.uchicago.edu/~pff/latent/

[5]Source code available at: www.mis.tu-darmstadt.de/code/

[6]Source code available at: www.eecs.berkeley.edu/~lbourdev/poselets/

The DPM model also relies on *pictorial structures* but differs in the prior imposed on the body parts. Here, a star shape prior is used, where all body parts are directly connected to the root part. Another difference is the interpretation of parts: While PS relies on manually labeled annotations, DPM automatically discovers the body parts that correspond to visually salient reoccurring structures in the training data. The configuration of body parts that maximizes Eq. (1) is found with max-product belief propagation. The entire model is trained in a purely discriminative fashion using the max-margin formalism. The appearance of each body part and the root part are trained with SVMs, while a deformation cost of part constellations is obtained with gradient descent. DPM is specifically optimized for detection. Since no part annotations are required, it can be trained on a significantly larger training set than PS, which requires such part annotations.

The final approach to part-based people detection considered in our experiments is the poselet-based detector (PBD) recently proposed by [5]. Instead of using the pictorial structures framework with a fixed number of parts, PBD relies on a large number of part detectors for diverse body regions denoted as "poselets", which have consistent appearance and correspond to similar 3D body configurations. Detections of different poselets are integrated using a probabilistic voting procedure resembling the implicit shape model [25] with weights learned using a max-margin framework [26]. Here every poselet votes for the location of the torso part, which in turn delimits the detection bounding box. Since the model is specifically designed to be robust to viewpoint and articulation changes, poselets often account for body regions that do not change significantly across articulation and viewpoint such as frontal faces, or correspond to frequently assumed body poses, such as legs of a standing person.

A. Proposed extensions

We propose two different kinds of extensions to the described vision based models: i) Since the different detectors focus on different aspects to be modeled, we propose to combine the complementary outputs of different detectors. ii) We introduce an extension that combines the vision based models with prior information obtained by the inertial sensors of the quadrotor.

Combining multiple models. The pictorial structures framework does not explicitly take the occlusion of body parts into account even though they frequently occur in our scenario. They happen due to complex poses in which some body parts are not visible, due to parts being outside of the view of the on-board camera, and due to miss-detections of some of the body parts from extreme foreshortening. In these cases the occluded body parts are fitted to spurious detections in the background, which results in a small probability of the overall configuration. In order to mitigate this problem we propose to combine the detection results of multiple models, each of which focuses on a different combination of body parts.

The DPM implementation used in our experiments is composed of two components, one upper-body and one full-body model. We extend the PS detector in a similar way and, complementary to a standard full-body detector, train an additional upper-body model, which is composed of torso, head, as well as upper and lower arms.

Fig. 3. Original image taken by quadrotor at the height of 1.77 meters (left) and ground plane projection (right). The shown rectangles correspond to ground truth annotations.

In order to fuse different models, we compute the posterior probability of each hypothesis k given the detection score d_k of model $\mathcal{M}$ as:

$$p(h_k|d_k,\mathcal{M})) = \frac{p(d_k|h_k,\mathcal{M})}{p(d_k|h_k,\mathcal{M}) + p(d_k|\neg h_k,\mathcal{M})}, \qquad (2)$$

where h_k is a Boolean variable corresponding to k-th hypothesis indicating whether it is correct or incorrect. $p(h_k|\mathcal{M})$ cancels as it is assumed to be uniform. The conditional distributions $p(d_k|h_k,\mathcal{M})$ and $p(d_k|\neg h_k,\mathcal{M})$ are assumed to be Gaussian, and fitted on a set of positive and negative detections.

The hypotheses of all models paired with the posterior probability are then cast forward to a joint non-maximum suppression step. Here, only the maximum scored detection is retained if several hypotheses overlap significantly. As the experiments demonstrate, this extension significantly improves the detection performance, especially on partially occluded people.

Scale prior based on UAV sensor measurements. The people detection methods discussed so far operate under the assumption that the camera position and depth for each image pixel are unknown. This implies that no prior information about the scale of the people in the image is available, and each model has to be exhaustively evaluated over all possible scales. However, in our scenario the quadrotor system is equipped with a calibrated camera and sensors capable of measuring the height and pitch angle of the vehicle. Combining these measurements allows to estimate the distance to the ground plane for each image pixel. Since our focus is on detecting people lying on the ground, this in turn provides an estimate of the scale of the person given an image position, subject to natural variation in people height and sensor noise. Similarly, knowing the position of the camera with respect to the ground plane we can back-project an image onto the ground plane taking both the homography transformation and image distortion into account [22]. An example of this projection is shown in Fig. 3. Note that while the scale of people differs in the original image, after back-projection it becomes approximately the same. Additionally, the camera calibration and back-projection enables the relation of the height of the detection bounding boxes measured in pixels

to the height of people measured in meters, which in turn allows to define a prior distribution on the bounding box height.

In the pictorial structures models, the posterior over configurations given by Eq. (1) contains the factor $p(\mathbf{l}_0)$ corresponding to the prior distribution on the position, scale and rotation of the root part $\mathbf{l}_0$, which is typically assumed to be uniform. When applying people detectors on back-projected images, we substitute this uniform prior with a Gaussian prior

$$p(\mathbf{l}_0) = \mathcal{N}(f(\mathbf{l}_0)|\mu_h, \sigma_h^2), \tag{3}$$

where $f(\mathbf{l}_0)$ is a linear transformation that converts the height of an hypothesis in pixels into metric units, $\mu_h = 0.8$ corresponds to an average upper body height of the person in meters, and $\sigma_h^2 = 0.1$. Note that this scale information is propagated to the other body parts through the body model.

As we show in the experimental section, not all models equally benefit from these priors. The PS model appears to be more precise in estimating the scale of people in an image, compared to DPM. While such precision is often not necessary when we are interested in detection only, it turned out to be beneficial when prior information about the expected scale of the person becomes available.

IV. Experiments

A. Experimental setup

Dataset. The test set used in the experiments contains 220 images collected in an indoor office environment under uncontrolled daylight illumination conditions. During data collection our quadrotor was flying at a height between approximately 1.5 and 2.5 meters, capturing the images with interval of approximately 1 second. The captured dataset contains 285 ground truth annotations of people. Several sample images from the dataset are shown in Fig. 1.

When recording the test set we aimed to "simulate" difficulties typical for a search and rescue scenario: note the large diversity in poses of people present in the dataset; also note that many people are only partially visible, either because some parts of the body appear outside of the image, or due to self-occlusion, or due to occluding objects present in the scene. Obviously, in an ideal world, we would have access to a real and representative dataset from a real search and rescue situation. Besides the practical issues of obtaining such a dataset it is also unclear what such a "representative" dataset would be. So in order to increase the realism and difficulty of our evaluation, we decided not to train any of the evaluated people detection methods specifically for this scenario, but rather relied on the training sets provided with the respective method. Therefore, we explicitly evaluate the generalization performance of these methods to our test set, while simulating difficulties typical for search and rescue scenarios.

Evaluation methodology. In our dataset we annotated upper bodies of all people visible to at least 50%. For the evaluation we use the same criterion as in [18], where a detection hypothesis is considered correct if the ratio of the intersection over the union of ground truth annotation and detection rectangles exceeded 0.25. We have chosen to define the detection task as detecting the upper body of the person, since all people detection methods considered in this paper are either designed to detect upper bodies [5], [18] or can be easily adapted to do so. For each of the methods included in our comparison we are using the implementations and trained models made publicly available by the authors.

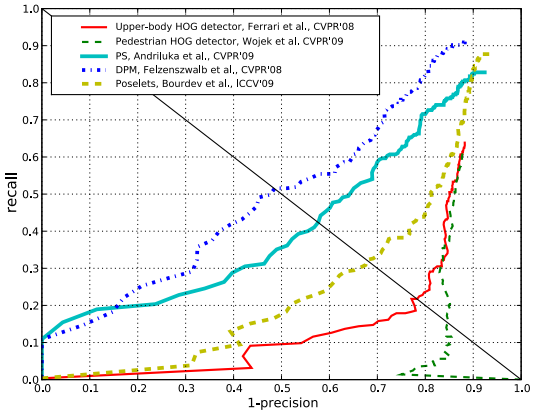

Fig. 4. Comparison of people detection methods

Our experiments consist of three parts: i) we compare different state-of-the-art methods of visual people detection, ii) we evaluate the importance of adding a scale prior to the model, and iii) we evaluate the performance of combinations of different detectors. For all of our experiments we report the equal error rate (EER) and show precision-recall curves.

B. Comparison of people detection methods

In our first experiment we evaluate the suitability of several recently proposed people detection methods for detecting articulated and partially occluded victims seen from our UAV.

The results are shown in Fig. 4 as recall-precision curves. Even though very common, the HOG based global template matching methods are not competitive in our setting. The HOG detector of Dalal and Triggs [9] achieves 15.1% EER. The conceptually similar upper body detector "HOG-upper" of [18] performs significantly better than the full body detector, but still achieves only 21.9% EER. Both HOG detectors do not perform well due to their monolithic structure, which does not take spatial variability in position of body parts into account. Several example detections of the full-body HOG detector are shown in Fig. 5 (first row). Note that while the HOG detector successfully copes with simple poses (e.g., bottom-left person in image (a)), it fails when body articulations vary significantly or when parts of the person become occluded as in images (b) and (c).

The poselet detector [5] achieves 32.0% EER. Several example detections are shown in the second row of Fig. 5. Compared to the monolithic HOG detectors, the poselet detector appears to be more robust to partial occlusions. Note the correct localization of partially occluded people in images (a), (b), and (c). However, the poselet detector appears to be challenged by poses in which characteristic parts of the upper

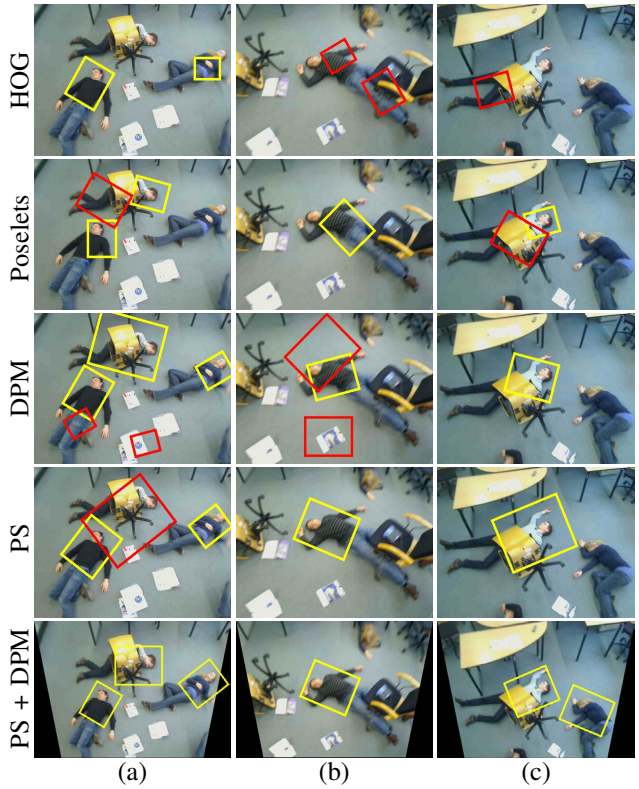

Fig. 5. Several examples of detections at EER obtained with a **full-body HOG detector [9]** (1-st row), the **poselet detector [5]** (2-nd row), the **DPM detector [15]** (3-rd row), the **full-body pictorial structures detector [1]** (4-th row) and the combined detector augmented with scale prior proposed in this paper (5-th row). True positive detections are plotted with yellow and false positives with red color.

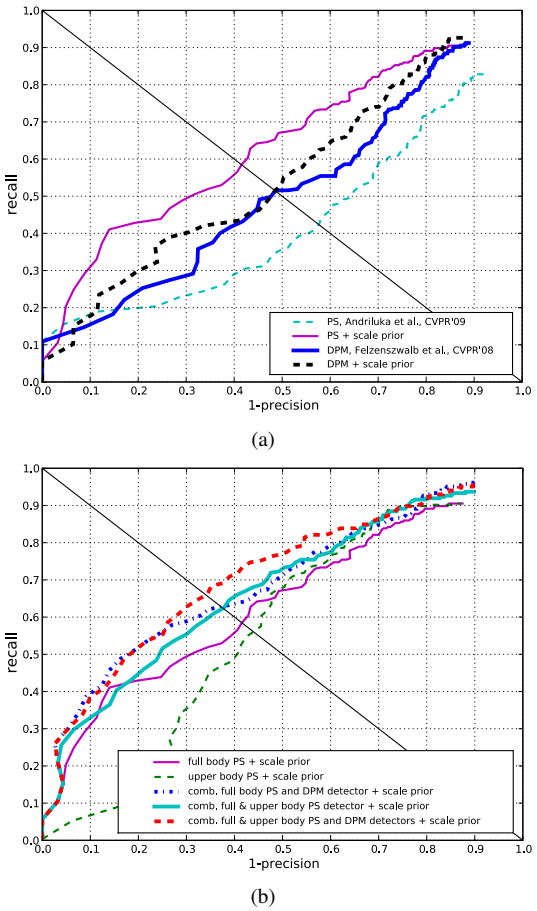

Fig. 6. Comparison of performance with and without scale prior (a), and evaluation of different model combinations (b).

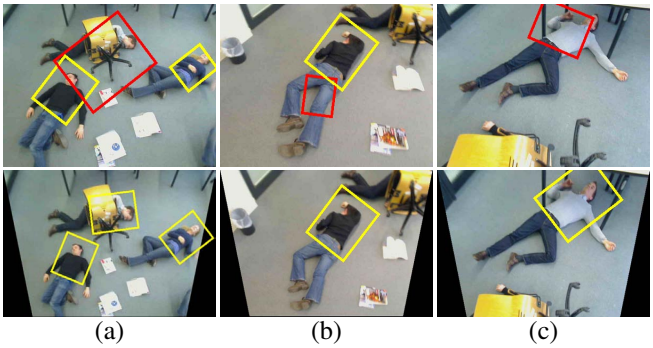

Fig. 7. Examples of people detection at EER obtained with the **full-body pictorial structures detector [1]** without scale prior (first row) and with scale prior (second row).

body are not visible, e.g., the rightmost person in Fig. 5(c). Additionally poselets seem to lack localization precision: Upper bodies are frequently localized with slight offsets from the correct position and scale, e.g. Fig. 5(b).

The two best performing detectors are both built on the pictorial structures framework. The PS detector [1] and the DPM detector [17] achieve 42.5% and 51.5% EER respectively. Note that this corresponds to a performance improvement over the monolithic model [9] by 27.4% EER and 36.4% EER. The difference in performance between the PS and DPM detectors is most likely due to a significantly larger number of images used to train the DPM detector and the fact that the DPM model internally combines 2 models corresponding to a full-body and an upper-body configuration, while the PS detector uses a full-body model only. We have found that, although the DPM model yields better detection performances, it is often less precise in localizing people compared to the PS detector (see Fig. 5 images (b) and (c)). Such behavior might be due to the discriminative training procedure employed in the DPM model, which is specifically optimized for detection. This procedure does not reward improvement in localization beyond the limit set by the bounding box matching criterion. In contrast, the PS model is specifically designed for localization and body part detection and uses generative learning to estimate parameters of pairwise part relationships.

C. *Integration of scale prior*

Even the two best performing models [1], [17] frequently suffer from false positive detections. One source of such false positives are detections at incorrect scales. An example of such false positives produced by the PS model is shown in the fourth row of Fig. 5(a). Note that these false positives frequently correspond to unreasonable sizes of the human body when back-projected into world coordinates. For example the false positive detection in the fourth row of Fig. 5(a) would correspond to a person with a height of approximately 3.5 meters. As described in Sec. III-A we can reduce the

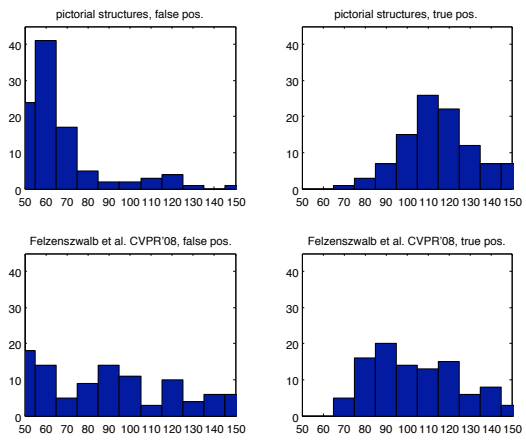

Fig. 8. Distribution of the height of false positive (left) and true positive (right) detections for PS [1] (top row) and DPM [17] (bottom row) detectors.

influence of this kind of false positives by extending the detectors with a prior distribution on the height of detected people. This is accomplished by first projecting images onto the ground plane and subsequently introducing a Gaussian prior using known relations between pixels and metric units.

We compare the influence of the scale prior on the detection results for the two best performing methods in our evaluation. The results are shown in Fig. 6(a). While both models benefit from the scale prior, the improvement for the PS model is almost 16% EER, and is more significant than the improvement for the DPM model, which improves only slightly overall without improving the EER. An insight into these different behaviors can be gained by examining the distribution of scales of the true and false positive detections in the images projected onto the ground plane. These distributions are shown in Fig. 8 for the PS and DPM methods. Note that the PS model distribution of true positives has a clear peak around 120 pixels, which roughly corresponds to the upper-body height of 85 cm. False positives on the other hand occur mostly at small scales. For the DPM model the height of true positives is distributed almost uniformly in the range between 80 and 120 pixels. The DPM model appears to be less precise in scale estimation, which however is not reflected in the recall precision curve in Fig. 4 due to the rather loose bounding box matching criteria. However, this imprecision turns out to be a handicap when information about the detection scale is available from other sources. Fig. 7 shows several examples of detections of the original PS model, and the PS model augmented with the scale prior. Note that in addition to removing false positives as in images (a) and (b), the back-projection to the ground plane removes effects of perspective distortions, which also improves detection results, as for example in image (c).

D. Combination of multiple detectors

Although both the PS and DPM detectors are built on the same pictorial structures framework, they differ significantly with respect to which parts are used in the model, which relationships between parts are considered and how the model parameters are learned from training data. The DPM model utilizes generic body parts that are automatically learned so that they are both discriminative and easy to localize in images. The DPM model appears to be more robust to occlusions since model parts are not fine tuned to detect particular parts of the body. This is in contrast to the PS model, which is designed to detect the actual body parts such as legs, torso, and head. As we found in our experiments, the PS model is superior to the DPM model in estimating the scale of a person, due to its more sophisticated body model enabling it to take advantage of a larger portion of the image evidence. In order to explore the complementarity of the PS and DPM detectors we derive a new detector based on their combination following the procedure described in Sec. III-A. In addition to the original DPM and full-body PS detectors we also train an upper-body PS detector. The results of this experiment are shown in Fig. 6(b). In isolation, the upper-body PS detector did not perform nearly as well as the full-body PS detector, however the combination of these two detectors improves the EER from 58% for the full-body PS detector to 62%. A similar performance improvement is achieved when combining the full-body PS and DPM detectors. The best results are obtained by the detector combining full-body PS, upper-body PS and DPM detectors, which achieves 66% EER.

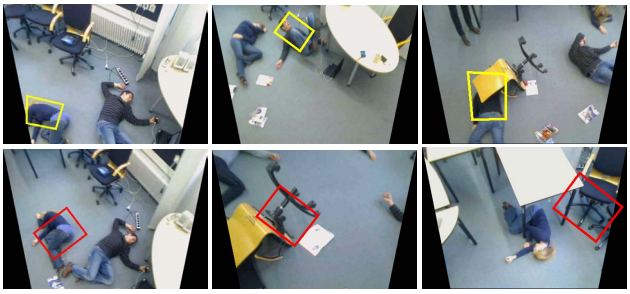

Fig. 10. Missing recall (top row) and false positive (at ERR) detections (bottom row) of the detector combining upper- and full-body PS, and DPM models.

Several examples of correct detections and false positives are shown in Fig. 5 (bottom row) and Fig. 9. Note that compared to previously proposed detectors, our improved detector is able to find people occluded by the armchair in Fig. 5(a) and the strongly articulated person in Fig. 5(c). Top row of Fig. 10 shows several examples of people not detected by our system. Note, that such missing detections correspond to people with either especially severe occlusions as in images (a) and (c) or particularly complex articulations as in image (b). Such complex cases in which only few body parts are visible, appear to be beyond the capabilities of state-of-the-art detection methods. The bottom row of Fig. 10 also shows a couple of false positives obtained by our system at EER. While some of them correspond to nearly correct detections as in image (a), the detector also occasionally fires on background structures as in images (b) and (c).

V. Conclusion

This paper evaluated the applicability of several state-of-the-art people detectors for victim detection from a UAV

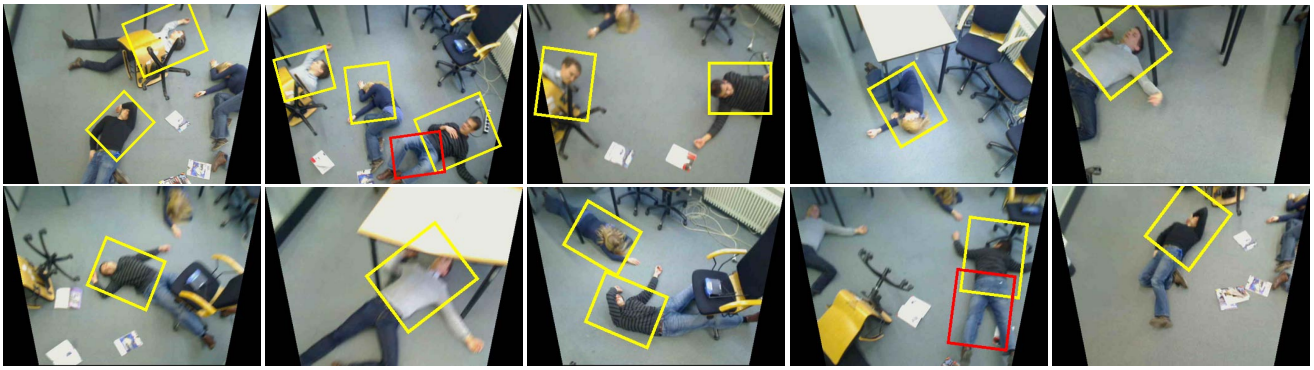

Fig. 9. Examples of detections at EER obtained with the detector combining upper- and full-body PS, and DPM models, and scale prior.

in a challenging search and rescue scenario. An important result of this comprehensive evaluation is that part-based models are better suited for victim detection than monolithic models, because they are able to represent variations in articulation and are robust to partial occlusions. As an extension to previous vision-based detectors we proposed to leverage complementary information of i) several detectors and ii) visual detectors and inertial sensor data of the UAV. Experimentally, we demonstrated that our extended framework substantially improved the detection performance, thus making a step towards autonomous victim detection in real world scenarios. We will make the collected images and the corresponding sensor measurements publicly available in order to foster further research on victim detection with UAVs.

VI. Acknowledgements

This work has been funded, in part, by GRK 1362 "Cooperative, Adaptive and Responsive Monitoring in Mixed-Mode Environments" of the German Research Foundation (DFG).

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. *CVPR 2009*.

[2] K. Arras, S. Grzonka, M. Luber, and W. Burgard. Eficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. *ICRA*, 2008.

[3] K. Arras, O. Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. *ICRA*, 2007.

[4] J. Y. Bouguet. Camera calibration toolbox for Matlab. www.vision.caltech.edu/bouguetj.

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. *ICCV 2009*.

[6] H. Bruyninckx. Open robot control software: The OROCOS project. *ICRA*, pages 2523–2528, 2001.

[7] A. Carballo, A. Ohya, and S. Yuta. Multiple people detection from a mobile robot using double layered laser range finders. *ICRA Workshop on People Detection and Tracking*, 2009.

[8] J. Cooper and M. Goodrich. Towards combining UAV and sensor operator roles in UAV-enabled visual search. *International Conference on Human Robot Interaction*, pages 351–358, 2008.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR 2005*.

[10] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *ECCV*, 2006.

[11] J. Davis and V. Sharma. Robust detection of people in thermal imagery. *ICPR*, 2004.

[12] P. Doherty and P. Rudol. A UAV search and rescue scenario with human body detection and geolocalization. *Australian Joint Conference on Artificial Intelligence*, 4830:1, 2007.

[13] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. *CVPR*, 2009.

[14] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. Moving Obstacle Detection in Highly Dynamic Scenes. *ICRA*, 2009.

[15] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2009.

[16] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61:55–79, Jan. 2007.

[17] P. F. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *CVPR 2008*.

[18] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. *CVPR 2008*.

[19] A. Fod, A. Howard, and M. Mataric. Laser-based people tracking. *ICRA*, 2002.

[20] G. Gate, A. Breheret, and F. Nashashibi. Centralized fusion for fast people detection in dense environment. *ICRA*, 2009.

[21] W. Green, K. Sevcik, and P. Oh. A competition to identify key challenges for unmanned aerial robots in near-earth environments. *ICAR*, 2005.

[22] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. *CVPR*, page 1106, Washington, DC, USA, 1997. IEEE Computer Society.

[23] G. Hoffmann, H. Huang, S. Waslander, and C. Tomlin. Quadrotor helicopter flight dynamics and control: Theory and experiment. *AIAA Guidance, Navigation and Control Conference*, 2007.

[24] A. Kleiner and R. Kuemmerle. Genetic MRF model optimization for real-time victim detection in Search and Rescue. *IROS*, 2007.

[25] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. *ECCV Workshop on statistical learning in computer vision*, pages 17–32, 2004.

[26] S. Maji and J. Malik. Object detection using a max-margin hough tranform. *CVPR 2009*.

[27] J. Meyer and A. Strobel. A flexible real-time control system for autonomous vehicles. *ISR / ROBOTIK*, 2010, to appear.

[28] R. Murphy. Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 34(2):138–153, 2004.

[29] K. Nordberg, P. Doherty, G. Farnebck, P.-E. Forssen, G. Granlund, A. Moe, and J. Wiklund. Vision for a uav helicopter. *Proceedings of IROS'02, Workshop on aerial robotics*, 2002.

[30] Q. Pham, L. Gond, J. Begard, N. Allezard, and P. Sayd. Real-time posture analysis in a crowd using thermal imaging. *CVPR*, 2007.

[31] D. Schulz, W. Burgard, D. Fox, and A. Cremers. People tracking with mobile robots using sample-based joint probabilistic data association. *IJRR*, 22(2), 2003.

[32] L. Spinello, R. Triebel, and R. Siegwart. Multimodal people detection and tracking in crowded scenes. *AAAI*, 2008.

[33] F. Suard, A. Rakotomamonjy, A. Benshrair, and A. Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. *IEEE Symposium on Intelligent Vehicle*, 2006.

[34] D. Titterton and J. Weston. *Strapdown inertial navigation technology*. American Institute of Aeronautics, Washington, DC, 2004.

[35] M. Villamizar, J. Scandaliaris, A. Sanfeliu, and J. Andrade-Cetto. Combining color-based invariant gradient detector with HoG descriptors for robust mage detection in scenes under cast shadows. *ICRA*, 2009.

[36] C. Wojek, S. Walk, and B. Schiele. Multi-cue onboard pedestrian detection. *CVPR 2009*.