How to improve the HOG detector in the UAV context

Paul Blondel, Alex Potelle, Claude Pegard, Rogelio Lozano

How to improve the HOG detector in the UAV context

P. Blondel * **Dr A. Potelle** * **Pr C. Pégard** * **Pr R. Lozano** **

* *MIS Laboratory, UPJV University, Amiens, France (e-mail: paul.blondel,alex.potelle,claude.pegard@u-picardie.fr)*
** *Heudiasyc CNRS Laboratory, Compiègne, France (e-mail: rogelio.lozano@uds-utc.fr)*

Abstract: The well known HOG (Histogram of Oriented Gradients) of Dalal and Triggs is commonly used for pedestrian detection from 2d moving embedded cameras (driving assistance) or static cameras (video surveillance). In this paper we show how to use and improve the HOG detector in the UAV context. In order to increase the elevation angular robustness we propose to use a more appropriate training dataset and sliding windows. We show results on synthetic images.

Keywords: HOG, UAV, human detection, computer vision, supervised training

1. INTRODUCTION

With the general lowering of UAVs' price and the recent progresses in this field, this technology is becoming more democratic for laboratories and companies of every size. UAVs are more and more used for various tasks. Nowadays they are currently considering using UAVs for searching and rescuing people or guarding specific areas such as nuclear plants or other sensitive areas. For these tasks embedded human detection algorithms are required in order to automatically detect people from the air.

1.1 Detection with background substraction

The detection of moving objects is obtained from the difference between the current frame and a reference frame, often called background image. The moving regions are analysed in order to classify the moving objects. The analysis can be done using a visual codebook (Zhou and Hoang, 2005), by using a contour shape matching (Toth and Aach, 2003) or by using other complementary information such as a depth-map and perform tests to reject objects not looking and not behaving like a human being (Xu and Kikuo, 2003). In complement to motion the thermal imagery is an interesting cue to identify human beings (Han and Bhanu, 2007). Hoewever, these methods are not suitable for human detection using a moving camera.

1.2 Visible-only detection

Human detection is also possible using the visible information of one camera. Gavrila and Giebel (2002) proposed to use a hierachy of human contour templates obtained after a training. This hierachy is used with the chamfer matching algorithm to detect people. More discriminative methods based on powerful descriptors were also developed. Descriptors permit to locally extract visual information. The collected information is compared to a general model of the object with a classifier. Papageoriou and Poggio (2000) were among the first to propose such a pipeline. They used wavelet descriptors, a sliding-window method to exhaustively scan the image and a SVM classifier. Many of nowadays' object detectors are still based on this approach. Viola and Jones (2001) based their work on the work of Papageoriou and Poggio (2000) and proposed the use of integral images and a cascade classifier to speed up the computation of Haar-like wavelets features and reach real-time performances for face detection. The Histogram of Oriented Gradients (HOG) detector (Dalal and Triggs, 2005) is an efficient people detector using a variant of the very well-known and quite efficient SIFT descriptor (Lowe, 1999). Visual information are extracted using SIFT-like descriptors over a sliding-window and compared to a general model using a linear SVM classifier trained on people images. This detector follows the classic pipeline. Zhu et al. (2006) proposed to use integral images and a cascade of HOGs to speed up the computation. Wang et al. (2009) proposed a more robust detector by using local binary patterns in complement with the HOG. The SIFT-like HOG descriptor is still one of the most descriminative descriptor for object detection.

1.3 Multiple information approach for detection

There are approaches combining multiple descriptors, image features and/or cues to increase the detection rate. Wojek and Schiele (2008) showed that combining HOG, Haar-like descriptors, shaplets and the shape context outperform the HOG detector alone. Dollár et al. (2009) proposed a mix between the detector of Viola and Jones (2001) and the HOG detector. This detector computes very simple rectangular features on integral images of several image channels : L,U,V, gradient magnitude and six "HOGs channels". A fast soft-cascade classifier is used to classify. Benenson et al. (2012) proposed another variant of the detector of Dollár et al. (2009) but running at

100Hz by both using stereo information and a multi-scale learning.

1.4 Multiple parts detection

The human body can be considered as a set of parts instead of one big monolithic part. Felzenszwalb and Huttenlocher (2005) proposed a method to detect people by parts and re-build human models from these parts by using a pictural structure reprensentation. Each part of the human model has to be separately learned and an incorrect labelling of human parts could decrease the performances of the detector (Felzenszwalb et al., 2010). Felzenszwalb et al. (2010) introduced the latent SVM classfier : during the training phase the most discriminative information are selected so that the final trained classifier be more robust (Felzenszwalb et al., 2010).

1.5 Existing work for detecting human from the air

Detecting human beings from a UAV is a tough task, and it becomes a tougher task if we take into account the various possible human poses. Most human detectors focus on detecting people in upright poses at nearby distances and from a more or less invariant view-point. The current two main applications of human detection are the security watching and the avoidance of pedestrians in the street. Nowadays little work has been done on detecting human from a UAV.

Gszczak et al. (2011) proposed to use both thermal and visible imagery to better detect people and cars, features extracted on thermal and visible imagery are fused together boosting the confidence level of detection. Indeed, the thermal camera is used for extracting Haar-like features while the optical camera is used for a contour shape analysis as a secondary confirmation to better confirm the detection. This method permits to detect upright people at about 160m distance, with a fixed camera pitch of minus 45 degrees and in real-time. This method doesn't seem suitable for detecting people closer to the UAV.

Rudol and Doherty (2008) also uses both thermal and visible imagery but in a pipeline way. They first identify high temperature regions of the thermal image and they reject the regions not fitting a specific ellipse. The corresponding regions are then analyzed in the visible using a relaxed Haar-like detector using the Leinhart extention. Upright and seated people can be detected with this method. However the more the UAV is close or at low altitude the more the thermal imagery become noisy and then tricky to analyse using only thresholding and ellipse fitting. Performances seem dependent on the distance.

Reilly et al. (2010) have a different approach. Their method is based on using shadows of human beings as a clue to localize human beings. The main problem with this technique is that we have to make strong assumptions on weather conditions and it also seems to be de-facto dependent on altitude.

In a different way Andriluka et al. (2010) evaluate various existing detection methods for detecting victims at nearby distances. According to this study, part-based detectors are better suited for victim detection from a UAV because they natively take into account the articulation of the human body. Part-based detection is also better suited for detecting from complex view-points. The authors propose to use complementary information using several detectors and inertial sensor data to reach better detection rate. However part-based detectors are not real-time (Felzenszwalb and Huttenlocher, 2005) and they are not very suitable for detecting people far from the camera.

1.6 Content of the paper

Automatically detecting people implies to take into account many different parameters. They are : the position and the orientation of the embedded camera according to the target, the distance, the variability of human poses, the illumination (because it can rapidly change and/or be different in the same image), the occlusion with objects of the environment, etc. In this paper we want to detect upright people in cleared areas and in a 10 to 80m distance range. This work is mainly about the management of the distance and the management of the position and the orientation of the camera according to the target.

This paper is a study of the HOG detector (Dalal and Triggs, 2005). In the first part, the general principle of the HOG detector is explained and the key configuration parameters are discussed. The UAV context is described and analysed. The second part is about how extend the boundaries of the detector with what improvements. The final part is about the analysis of the results obtained with these improvements.

2. STUDY OF THE HOG DETECTOR IN A UAV CONTEXT

2.1 The HOG detector

How it works The detection is performed as it follows : the image is exhaustively scanned by a sliding detection window of a specific size and ratio as showed in figure 1b. An object is detected if the combination of all the histograms computed within this detection window match a general model of the object class. In order to detect objects of different sizes we build an image pyramid from the original input image and we scan all the levels as showed in figure 1a. The configuration of this image pyramid is directly related to the expected sizes of the objects we are looking for.

For each detection window the histograms are computed in a very specific manner. The detection window is composed of overlapped blocks as showed in figure 1c (in blue). A block is composed of a certain number of cells. For each cell we compute an histogram of the oriented gradients. Typically a block is composed of four squared cells. The histogram is divided by bins, typically nine bins from 0deg to 180deg as recommanded by Dalal and Triggs (2005). All the histograms of the blocks are finally locally normalized using the L2-Norm or the L2-Hys Norm.

The data computed within the sliding detection window is compared to a general model. The general model of the object class is built using a SVM classifier trained using the appropriated training images (positive and negative case images).

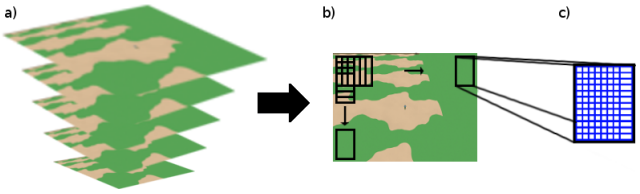

Fig. 1. Image pyramid, scanning with the sliding window, computation of the histograms and normalization for each overlapped block (in blue)

The image pyramid An image pyramid is required for finding objects of different sizes. Building the right image pyramid is very important. Three parameters are required to build an image pyramid : a number of levels or a scale factor, the minimum and the maximum scale. If the setting of the parameters is maladjusted then objects could be missed. The default implementation of the HOG detector uses a scale factor of 1.05 between following levels.

The detection window The ratio of the detection window is important, a vertical one-half ratio is usually chosen in case we want to detect upright people in a pedestrian-view scenario. We could think of using an horizontal one-half ratio to detect cars but this is obviously dependent on the view-point. The ratio of the detection window should change with the view-point to better match the shape of the object, yet it depends on the nature of the object. Changing the ratio of the detection window often requires to change the block configuration.

The training images Detection performances could be improved by a more judicious choice of the training images. Indeed, it is possible to chose the negative training images according to the environment in order to better train the classifier for the specific hard cases we can encounter in this environment. Moreover, choosing better positive training images can also improves the detection rate by reinforcing the general object model. The training image dataset should be revelant to the UAV context.

2.2 The UAV Context

Unconventional view-points UAVs move in a 3d world. A drone's camera undergoes rolling, pitching, heading or a combination of all and this makes the detection more complex. Even with a camera stabilizer there will always be an important elevation angle between the scanned area and the drone. Besides as people can be located at different places of the scanned area, the elevation angles from the human beings to the camera can be very different. To detect people from the air the detector must be capable of dealing with angles, and especially with the elevation angle.

Wide distance ranges People can be far or close to the drone. The distance range can be very important and then requires many more scanning with many more scales (big image pyramid). The bigger the image pyramid is the more the detection is timeconsuming. Besides, people can be missed with a too spaced image pyramid. A naïve approach would be to focus the detection process on a specific distance range and the movement of the UAV would make possible an exhaustive scanning of the area.

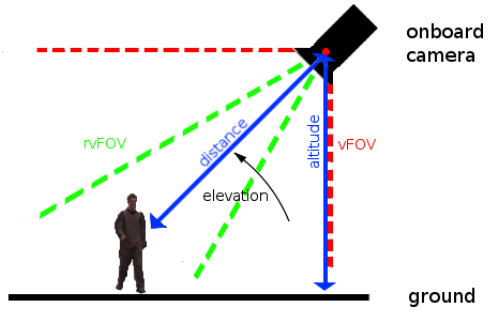

Fig. 2. Using elevation and distance for building image pyramid

All we need to know is the distance along the camera axis from the camera to half the average size of people. The distance can be obtained using the elevation angle and the altitude of the drone minus half the average size of people. The principle is illustrated in figure 2, vFOV means vertical field of view and rvFOV means restricted vertical field of view. Of course, we have to make the assumption the ground is flat or almost flat. If we know the average expected size of human beings at a certain distance then we can deduce the average expected size of human beings at any distances and thus build a dense image pyramid around this scale. A wider image pyramid can be built by repeating this principle on a restricted vertical field of view as showed in figure 2.

Changing weather conditions UAVs are outdoor robots, they are subject to weather conditions. The detection method should be robust to illumination changes and not relies on human shadows such as (Reilly et al., 2010) did. The HOG detector is natively quite robust to illumination changes because of the local block normalization but a multi-cues approach could help make the detection even more robust to this.

3. EXTEND THE BOUNDARIES OF THE DETECTOR

3.1 Important aspects to improve

Dealing with angles To be really usable in a UAV context the HOG detector must be robust to elevation angles between people in the area and the drone's camera, as it is showed in figure 3. This angle cannot be compensated.

Attenuate the impact of people distance The distance has too drawbacks : the more people are far from the camera the more it is possible to miss them during the detection and it takes more computation time. Indeed, detecting smaller people requires to build an image pyramid with bigger image scales, thus more data has to be computed and compared. The classic HOG implementation is not appropriate when people are far from the camera, as in a UAV context.

3.2 The datasets

The synthetic images were generated using POV-Ray. The human models were generated with Makehuman and the different poses mimiced using Blender. Eighteen different human models were generated. Half of the models were

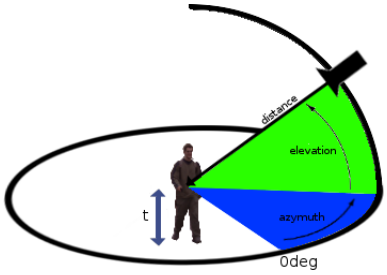

Fig. 3. Camera position for virtual image training

Fig. 4. GMVST image examples

Fig. 5. GMVST2 image examples

male, the other half were female. Sub-models were generated by three times slightly changing the pose of each model and twice for the last model. At the end there were sixty different human models for the training images. The three different poses were : people walking, people making distress signs and people in relaxed pose. Tweenty other models were generated for the testing images. They were generated the same way but using different human models.

GMVST : Generalized multi-view synthetic training dataset[1] The positive training images were taken around the human models with a 10deg interval. The elevation angle changes from 0deg to 90deg with an interval of 10deg in order to cover all the views all around the human models. Each time an environment was randomly chosen from five different ones : desert-like sand, algae, sand + algae, grass and sand + algae. These images were finally cropped around the human models with a fixed padding of about 10 pixels and resized to 64x128. The negative training images were generated the same way but without human models. Several imagets of 64x128 were randomly extracted from each of these images to get the negative training images. Examples of positive training images are showed in figure 4. For illustration purpose 4 both the elevation and the azymuth are changing in figure 4.

GMVST2 : GMVST for multiple detection window sizes[1] The training images were generated the same way as explained for the GMVST images. However here the size of the training images changes with respect to the elevation angle. The size is 64x128 when the elevation angle in the scene is below or equal to 40deg. The size is 64x112 when the elevation angle is bigger than 40deg and smaller to 70deg. The size is 64x64 when the elevation angle is bigger than or equal to 70deg. Examples of positive training images are showed in figure 5.

Test dataset[1] The tweenty test models were used for this dataset. The distance from the virtual camera to the model changed from 10m to 80m with a step of 10m (after metric calibration). The elevation from the human models

[1] http://mis.u-picardie.fr/ p-blondel/papers/data

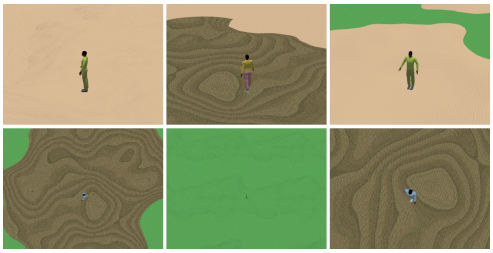

Fig. 6. Test image examples

to the camera changed nine times from 0deg to 90deg with a step of 10deg. The azmyuth angle was randomly chosen between 0deg and 360deg and an environment was randomly picked up among the five ones described above. Examples of test images are showed in figure 6.

INRIA training dataset[2] The INRIA training dataset is composed of real images. These pictures were taken in very different places so that the trained detector be very robust to the environment. This dataset contains an important number of human pictures.

3.3 Tests

Detector implementation and hardware The HOG detector was implemented as described and advised by Dalal and Triggs (2005). That is with 9 bin histograms, a three-quarter overlapping of the blocks and four cells by blocks. However the L2-Norm was prefered for the local block normalization. Indeed, the training was much faster using this norm, besides the gain obtained using the L2-Hys Norm instead of the L2-Norm seems minor (Dalal and Triggs, 2005). The tests were performed on a computer with a Intel i7 2.20GHz processor and having 8Go of memory.

Test 1 : Better image pyramid for speed By default the HOG detector as described by Dalal and Triggs (2005) uses a 64x128 sliding detection window to scan all the levels of the image pyramid. This window size is more appropriate when people are not that far away from the camera. People are more likely to be far from the camera in a UAV context. And looking for small or far objects requires to upsample a lot the images of the image pyramid with the default window size, and this is costly. This test is about speeding up the detector without degrading the performances by using a smaller sliding detection window. The test was performed on a four times downsampled INRIA test dataset. Three different trainings were done using differently sized INRIA training images. During the test the corresponding training was selected according to the sliding detection window size. The performances of three sliding detection window sizes were compared, with the following configurations : 64x128 window with sixty-four pixels by cell, 48x96 window with thirty-six pixels by cell and 32x64 window with sixteen pixels by cell. The number of pixels by cell was chosen so that the final result vectors have the same dimension whatever the window size in order to allow in theory the same data diversity. In Table 1 are exposed the three image pyramid configurations. Each detector was bootstrapped once by following the recommandations available on the INRIA website[2]. To

[2] http://pascal.inrialpes.fr/data/human/

be sure the boostrapping was proceeded in a similar way for the three detectors, the negative images used for the bootstrapping were resized by a three-quarter factor for the 48x96 detector and by a one-half factor for the 32x64 detector.

Table 1. Image pyramid configurations

detector	nscales	minscale	maxscale
64x128	64	0.6	4.26
48x96	64	0.45	3.2
32x64	64	0.3	2.13

Test 2 : Other window ratios and trainings for angular robustness The purpose of this test is the robustness to the elevation angle. The HOG detector has been trained with three different training datasets : the INRIA dataset, the GMVST dataset, and the GMVST2 dataset. The sliding detection window changes with respect to the elevation when using the GMVST2 training. The window size is 64x128 with 64 pixels by cell for an elevation angle below or equal to 40deg, the window size is 64x112 with 36 pixels by cells for an elevation angle above 40 and below 70 deg and the window size is 64x64 with 16 pixels by cell for an elevation angle above or equal to 70deg. The number of pixels by cell for the 64x112 and 64x64 windows were chosen so that the number of blocks be bigger than the number of blocks of the default HOG configuration. Thus, the dimension of the final result vectors are never smaller than the dimension of the vector obtained using the 64x128 detection window. The image pyramid was built for the three cases as proposed in part 2.2. This test was performed on the test dataset.

4. RESULTS

4.1 Test 1 : Better image pyramid for speed

The ROC curves showed in figure 8 were generated following Dollar's recommendations (Dollar et al., 2009) : the miss rate is plotted against the false positives per image (FPPI). As Dollar et al. (2009) we decided to use the miss rate at 1 FPPI as a common reference point to compare the detectors. As it is showed in figure 7, the performances of the three detectors are quite similar in this case. Using a more appropriatly sized sliding detection window does not decrease the performances so much if the number of blocks is the same and the number of pixels by cell is superior than the number of bins. We could think of using different sliding detection windows for different distances and thus easily reduce the computation time while keeping relatively equivalent performances.

4.2 Test 2 : Other window ratios and trainings for angular robustness

Results obtained with the INRIA dataset[2] The average detection rate falls significantly for an elevation angle of about 50deg and whatever the distance (figure 8a). The rate is about zero for an elevation angle bigger than 80deg. The time spent for the computation increases uniformly with the distance, reaching about 40 seconds by image for a distance of 80m (figure 8b). The FPPI number increases uniformly with the distance and reaches about 20 to 30 for a distance of 80m (figure 8c).

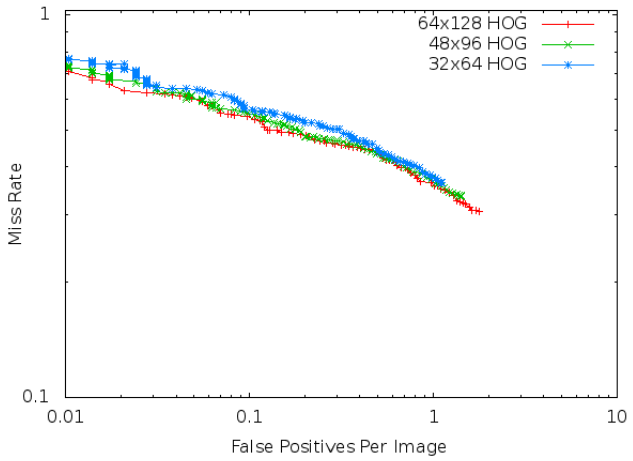

Fig. 7. Performances using different detection window sizes

Results obtained with the GMVST dataset[2] The average detection rate is almost the same whatever the elevation angle and the distance (figure 9a). People are detected in every configuration. The time spent for the computation increases more with this training dataset because the non-maximum suppression algorithm has more detections to treat and because the number of false positives is bigger (figure 9b and 9c). More complex datasets such as the INRIA dataset permits to train a detector quite robust to the surrounding environment because of the diversity of the INRIA negative training images.

Results obtained with the GMVST2 dataset[2] The average detection rate is almost the same although it slightly falls for a distances bigger than 70m. Nevertheless people are detected whatever the elevation angle and the distance (figure 10a). This time the FPPI number and the time spent on computation increase both with respect to the elevation and the distance (figure 10b and 10c). The 64x112 and 64x64 detection windows have more blocks and this slows down the classification. The FPPI number increases because the algorithm analyses smaller surfaces.

Synthesis Reducing the size of the sliding detection window does not decrease so much the performances and has the effect of reducing the size of the image pyramid levels and lighten the computation. Keeping the same block configuration is important. The block configuration was 7x15 and the number of pixels by cell was always bigger than the number of bins. We also pointed out that classic datasets are not appropriate in a UAV context. With multi-view datasets we better deal with the elevation angle and detect people in many angular configurations. Adujsting the size and the ratio of the detection window with respect to the elevation is costly and it increases the FPPI number. This is partially due to the changing number of blocks. Still, this last solution seems interesting for cluttered environments.

5. CONCLUSION

This work shows we can use the HOG detector to perform human detection from a UAV. The results show that adjusting the detection window with respect to the distance make senses. The results also show that multi-

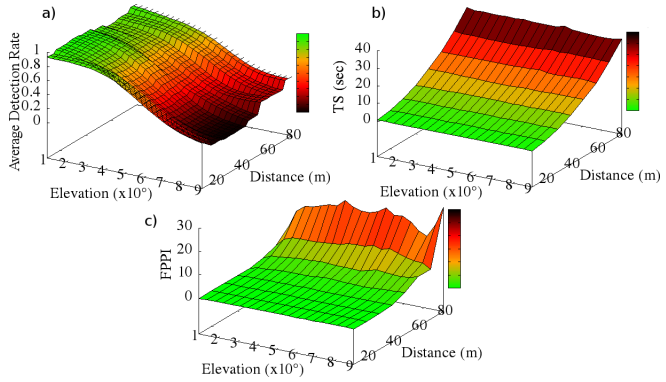

Fig. 8. Results obtained with the INRIA dataset [2]

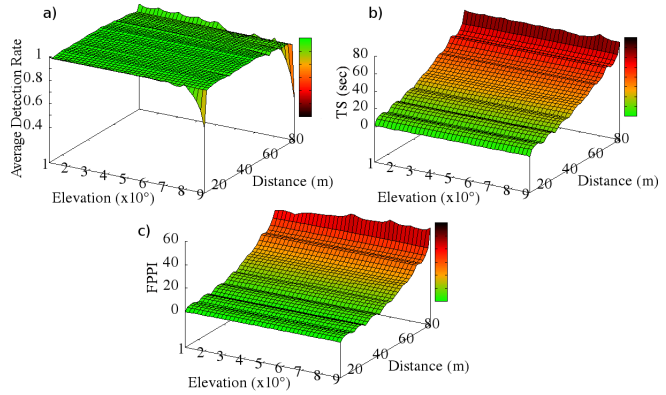

Fig. 9. Results obtained with the GMVST dataset [2]

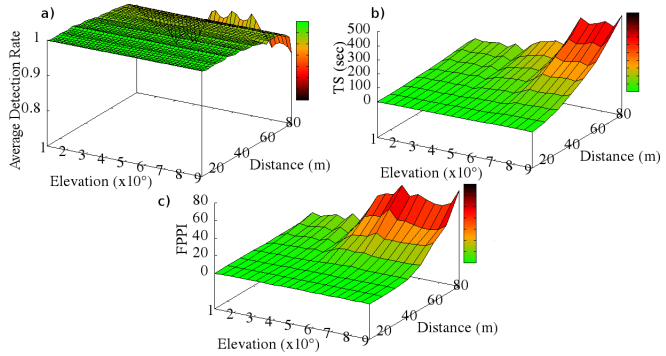

Fig. 10. Results obtained with the GMVST2 dataset [2]

view training datasets extend the angular robustness to the elevation angle. Similar tests have to be performed on real images : the average detection rates should be smaller and FPPI numbers bigger. This study mainly focuses on the detection of upright people. The next objective is to detect people with complex poses.

REFERENCES

Andriluka, M., Schnitzspan, P., Meyer, J., Kohlbrecher, S., Petersen, K., von Stryk, O., Roth, S., and Schiele, B. (2010). Vision based victim detection from unmanned aerial vehicles. In *Conference on Intelligent Robots and Systems (IROS)*, 1740–1747.

Benenson, R., Mathias, M., Timofte, R., and Gool, L.V. (2012). Pedestrian detection at 100 frames per second. In *Computer Vision and Pattern Recognition*, 2903–2910.

Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *Conference on Computer Vision and Pattern Recognition*, 886–893.

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2009). Pedestrian detection: A benchmark. In *Conference on Computer Vision and Pattern Recognition*, 304–311.

Dollár, P., Tu, Z., Perona, P., and Belongie, S. (2009). Integral Channel Features. In *Proceedings of the British Machine Vision Conference*, 91.1–91.11.

Felzenszwalb, P.F., Girshick, R.B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Transactions on pattern analysis and machine intelligence*, 32, 1627–45.

Felzenszwalb, P.F. and Huttenlocher, D.P. (2005). Pictorial Structures for Object Recognition. *International Journal of Computer Vision*, 61, 55–79.

Gavrila, D. and Giebel, J. (2002). Shape-based pedestrian detection and tracking. *Intelligent Vehicle Symposium*, 1, 8–14.

Gszczak, A., Breckon, T.P., and Han, J. (2011). Real-time People and Vehicle Detection from UAV Imagery. In *Intelligent Robots and Computer Vision*, 8–11.

Han, J. and Bhanu, B. (2007). Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 1771–1784.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *International Conference on Computer Vision - Volume 2*, 1150–1157.

Papageoriou, C. and Poggio, T. (2000). A Trainable System for Object Detection. *International Journal of Computer Vision*, 38, 15–33.

Reilly, V., Solmaz, B., and Shah, M. (2010). Geometric constraints for human detection in aerial imagery. In *European conference on Computer vision: Part VI*, 252–265.

Rudol, P. and Doherty, P. (2008). Human Body Detection and Geolocalization for UAV Search and Rescue Missions Using Color and Thermal Imagery. In *Aerospace Conference*.

Toth, D. and Aach, T. (2003). Detection and recognition of moving objects using statistical motion detection and Fourier descriptors. In *International Conference on Image Analysis and Processing*, 430–435.

Viola, P. and Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In *Computer Vision and Pattern Recognition*.

Wang, X., Han, T.X., and Yan, S. (2009). An HOG-LBP human detector with partial occlusion handling. In *International Conference on Computer Vision*, 32–39.

Wojek, C. and Schiele, B. (2008). A Performance Evaluation of Single and Multi-feature People Detection. *Pattern Recognition, 30th DAGM Symposium*, 82–91.

Xu, F. and Kikuo, F. (2003). Human detection using depth and gray images. In *Advanced Video and Signal Based Surveillance*, 115–121.

Zhou, J. and Hoang, J. (2005). Real Time Robust Human Detection and Tracking System. In *Computer Vision and Pattern Recognition*, 149.

Zhu, Q., Avidan, S., Yeh, M.c., and Cheng, K.t. (2006). Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In *Computer Vision and Pattern Recognition - Volume 2*, 1491–1498.