Human Detection in Uncluttered Environments: from Ground to UAV View

Paul Blondel, Alex Potelle, Claude Pégard, Rogelio Lozano

Human Detection in Uncluttered Environments: from Ground to UAV View

Paul Blondel
Université Picardie Jules-Vernes
80000 Amiens, France
Email: paul.blondel@u-picardie.fr

Alex Potelle and Claude Pégard
Université Picardie Jules-Vernes
80000 Amiens, France
Email: firstname.lastname@u-picardie.fr

Rogelio Lozano
Université Technologique de Compiègne
60200 Compiègne, France
Email: rogelio.lozano@hds.utc.fr

***Abstract*—Nowadays pedestrian detectors are fast, scale-robust and quite efficient. Embedded within a UAV such a detector would open new possibilities. In this paper the very well known HOG detector is adapted for UAV use and a new kind of training dataset is proposed in order to increase the detector's angular robustness. A more appropriate set of detection windows, together with a new detection pipeline, is proposed in order to reduce the search space and consequently reduce the computation time. Tests conducted using the improved detector show significantly better results on aerial images.**

I. Introduction

Given the continuous fall in UAV prices and technological advances in this field, UAVs are becoming more and more accessible to laboratories and companies of every size. UAVs are more and more used for various tasks. Nowadays they are currently considering using UAVs for search and rescue missions, or monitoring specific areas such as nuclear plants or other sensitive areas. These tasks require embedded human detection algorithms to automatically detect people from the air.

A. Existing work in pedestrian-view detection

1) Detection with background substraction: The detection of moving regions is obtained by the difference between the current frame and a reference frame, often called background image. These regions are analysed in order to classify the moving objects. The analysis can be done, for instance, with the help of a visual codebook [1] or by using contour shape matching [2]. A more powerful classification of the moving regions can be performed using a monolithic or a multi-parts human detector. Background substraction is not suitable when the camera itself is also moving.

2) Monolithic detection: Monolithic detectors look for monolithic parts of the image that look like people. Gavrila et al proposed to use a hierachy of human contour templates obtained using training [3]. This hierachy, used together with the chamfer matching algorithm, permits the detection of people in images. But more discriminative methods based on powerful descriptors have also been developed. The visual information is locally extracted and collected. Finally the information is compared to a general model of people with a classification algorithm. Papageroriou et al were among the first to propose this pipeline [4]. They used wavelet descriptors, a sliding-window method to exhaustively scan the image and a SVM classifier. Many of current detectors are still based on this approach. Viola et al based their work on the work of Papageoriou et al [4]. They used integral images and a cascade classifier to speed up the computation of the Haar-like wavelet features and reach real-time performance for face detection [5]. The Histogram of Oriented Gradients (HOG) detector of Dalal and Triggs [6] is an efficient human detector using a variant of the very well-known and quite efficient SIFT descriptor [7]. Visual information is extracted using SIFT-like descriptors over a sliding-window. All the information is classified using a linear SVM classifier trained on images of people. The SIFT-like HOG descriptor still remains very competitive for object detection.

Some detectors combine multiple descriptors, image features and/or information sources to increase the detection rate. Wojek et al showed that combining HOG, Haar-like descriptors, shaplets and the shape context outperform the HOG detector alone [8]. Dollar et al proposed a mix between Viola et al's detector and the HOG detector [9]. This detector computes simple rectangular features on integral images of different channels: L,U,V, gradient magnitude and six "HOG channels". The classification is performed using a fast soft-cascade classifier.

3) Multiple parts detection: Instead of considering the human body as one monolithic part, some detectors consider it as a set of parts. Felzsenzwald et al proposed a method to detect people by fragments and re-build human models by using a pictural structure reprensentation [10]. Each part of the human model is separately learned. An incorrect labelling of the fragments could decrease the performance of the detector [11]. That is why Felzenszalb et al introduced a detector using a new classifier: the latent SVM classifier [11]. With this classifier the most discriminative information is selected during the training to produce a more robust detection.

B. Existing work for detecting people from a UAV

Detecting people is difficult and it becomes more difficult in a UAV context. Most human detectors focus on detecting upright people at nearby distances and from a more or less invariant viewpoint. The current two main applications of human detection is the security monitoring and the driving assistance. Until now little work has been done on detecting humans from a UAV. Unlike a pedestrian view, an UAV view is more complex to manage because the drone undergoes pitching and rolling rotations. People are also on average further from the camera in this context.

Gaszczak et al proposed to use both thermal and visible imagery to better detect people and vehicles [12]. Features extracted on thermal and visible imagery are fused together to boost the confidence level of detection. 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 a distance of about 160m using a fixed camera pitch rotation of minus 45 degrees and in real-time. This method does not seem flexible enough for detecting people closer to the UAV.

Rudol et al also use thermal and visible imagery but in a pipeline way [13]. They first identify high temperature regions from the thermal image and they reject the regions not fitting a specific ellipse. The corresponding regions are then analyzed in the visible spectrum using a relaxed Haar-like detector. Upright and seated people can be detected with this method. However, the thermal imagery can easily become very tricky to analyze with this method when the drone is too close and the information becomes too noisy.

Reilly et al have a different approach [14]. They use people's shadows as a key clue to detect and localize people. But strong assumptions on weather conditions have to be made with this technique.

Andriluka et al evaluated various detection methods for detecting victims at nearby distances [15]. They showed part-based detectors are better suited for victim detection from a UAV because they natively take into account the articulation of the human body. The authors propose the use of complementary information using several detectors and inertial sensor data to obtain a better detection rate. However, part-based detection is a slow process [10] and this seems not suitable for detecting people too far from the camera.

C. Paper content

Different parameters have to be considered to automatically detect people from a UAV: the position and the orientation of the embedded camera in relation to the target, the distance, the variability of human poses, the illumination, occultation, etc. The purpose of this paper is to show that it is possible to easily adapt a pedestrian-view human detector for UAV-view human detection. The aim is the detection of upright people located between 10 and 40m from the camera in a fast and robust manner and in uncluttered environments. This work provides solutions to manage the distance, the search space and the orientation of the target in relation to the camera.

It was chosen to adapt one of the most well-known pedestrian detectors: the HOG detector of Dalal and Triggs [6]. The section II describes the HOG detector and its key configuration parameters. The UAV context is also discussed. Section III deals with ways to adapt the detector to the UAV scenario. In section IV experimental results are presented and discussed.

II. Study of the HOG detector in a UAV context

A. The HOG detector

1) How it works: The input image is exhaustively scanned by a sliding detection window of a specific size and ratio as shown in Fig.1 b. An object is detected if the combination of all the histograms computed within this detection window matches a general model of the object class. In order to detect objects of different sizes an image pyramid is built from the original input image and all the levels scanned as showed in Fig.1 a. The configuration of this image pyramid is directly related to the expected sizes of searched objects.

For each detection window the histograms are computed in a very specific manner. The detection window is composed of overlapped blocks as shown in Fig.1 c (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 into bins, typically nine bins from 0deg to 180deg as recommended by Dalal and Triggs [6]. At the end, all the histograms of the blocks are 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).

2) The image pyramid: An image pyramid is required to find 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. Objects can be missed if these parameters are maladjusted.

3) The detection window: The ratio of the detection window is important, a vertical one-half ratio is usually chosen to detect upright people in a pedestrian-view scenario. Changing the ratio with the viewpoint to better match the shape of the object could be considered, but changing the ratio of the detection window often requires changing the block configuration. Changing the block configuration tends to change the performance as well [6].

4) The training dataset: A more judicious choice of the training images can improve the detection performance. Choosing the negative training images according to the environment improves the performance of the classifier because environment specific hard cases are learned. Better positive training images improve the detection performance as well because it reinforces the general object model.

B. The UAV context

1) Unconventional camera angles: UAVs move in a 3D world. A UAV camera undergoes rolling, pitching, heading or a combination of all: it complexifies the detection. Camera

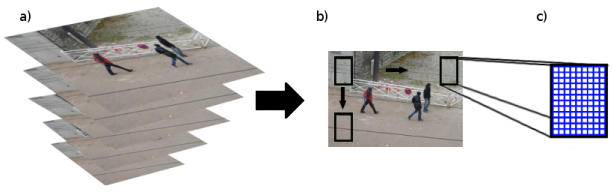

Fig. 1. a) image pyramid, b) scanning with the sliding window, c) computation of the histograms and normalization for each overlapped block (in blue).

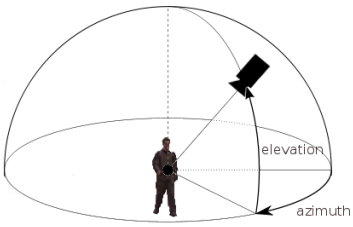

Fig. 2. Azimuth and elevation angles.

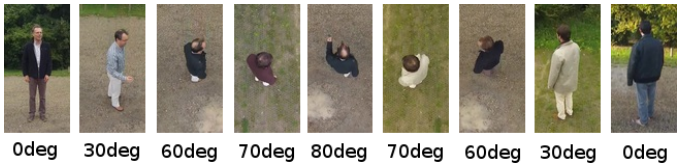

Fig. 3. Examples of GMVRT-V1 positive samples.

stabilizers do not solve all the problems: there will still be a great elevation angle between the ground and the UAV's camera. It can be summarized as follows: the rolling angle tends to rotate the shape of people and the elevation angle tends to shrink the shape of people (Fig.2, 3 and 4). In addition, as the elevation angle increases, perspective effect tends to rotate people far from the camera axis (horizontally) ; for example people in the left side of Fig.4 a and Fig.4 d.

2) Wide distance ranges: The greater the distance range the more scale scans are required. This increases the computation time. An image pyramid with less levels can be deduced using geometric knowledge of the scene when available. This is not the case in this work.

3) Changing weather conditions: UAVs are subject to weather conditions because they are outdoor robots. The detection should be robust to illumination changes. Fortunately the HOG detector is natively quite robust to this because of the local block normalization. But additional cues could be used in order to increase its robustness.

III. Adapt the detector to the UAV-view

A. A more appropriate dataset

Blondel et al showed that the robustness to the elevation angle can be improved on synthetic images by a multi-view training at different elevations [16]. When the elevation angle is greater than 45degrees significantly better results were obtained for detecting 3d models. We propose to extend and to evalute this approach in a real case scenario. A multi-view training dataset called GMVRT-V1 (Fig.3) and an aerial test dataset (Fig.4), for testing the detector performance, have been built.

1) GMVRT-V1 : Generalized multi-view real training dataset (version 1)[1]*:* Six different models were used for obtaining the positive samples. They were asked to mimic three different poses (relaxed, walking and making distress signs). They were also asked to change their clothes once. The acquisitions of the positive samples were made in the following manner: a GoPro Hero 3 camera was attached to the top of a triangle formed by two rods. Models were asked to stay below the camera and two others people synchronously rotated the triangle in an arc of a circle from 90 degrees elevation to 0 degree elevation for each type of model pose and orientation. The positive samples were taken at different elevation angles. The negative samples are images taken from natural and uncluttered environments. The barrel distortion effect has been corrected in each image.

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

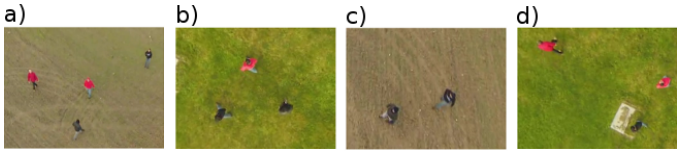

Fig. 4. Examples of images from the aerial test dataset.

2) Real images test dataset [1]*:* We took images of people with different natural backgrounds in an uncluttered environment. We made sure that the pitching angle of the camera was greater than 40 degrees. Aquisitions were made using a GoPro Hero 3 and images have been corrected to remove the barrel distortion effect.

B. Smaller detection windows for detecting smaller objects

1) Principle: The standard HOG uses a 64x128 window with 64 pixel cells to scan all the levels of the image pyramid [6]. This size is not suitable in a UAV scenario because people are more likely to be far from the camera. Bigger levels are required to look for far off objects/people and this is time-consuming. Smaller detection windows would partially help resolve this problem. But this should not alter the detection performance. A simple design is proposed to ensure that: one of the Dalal and Triggs' block configuration (2x2 cell blocks) with a number of pixels by cell always greater than the number of bins (in this case, 9). It gives three possible window configurations: 64x128 with 64 pixel cells, 48x96 with 36 pixel cells and 32x64 window with 16 pixel cells. The window configuration (and thus the associated classifier) could be dynamically switched from one configuration to another in relation to the scanning depth.

2) Complexity analysis: Taking a CCD camera with a 1/3" sensor, a 640x480 image resolution and a 8mm lens. As a first step it is considered that all the pixels of the level are first classified into the histogram bins. In a second step the block normalization is performed inside the scanning detection window. The two steps are repeated for each level. We consider a distance range from 10m to 40m with an average person size of 1m70. The scale range is from 0.53 to 1.06 for the 32x64 window, from 0.8 to 1.6 for the 49x96 window and from 1.06 to 2.13 for the 64x128 window.

The number of pixel classifications for one level (first step) is given by equation 1 (s is the scale):

$$Complexity(s) = totalSurface \times s = 640 \times 480 \times s \quad (1)$$

Figure 5 shows the complexity for a given level scale s. The total number of pixel classifications performed between two levels is given by the area under the curve. However, in practice, only a discrete approximation of the area is true.

For the 32x64 and 48x96 windows, the area under the curve is respectively 4.06 times and 1.77 times smaller than the one of the 64x128 window (Fig.5). In conclusion, using a window size of 32x64 is faster.

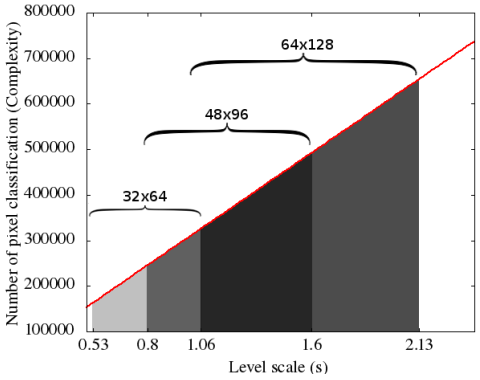

Fig. 5. In red : the growth of the complexity. From scale 0.53 to 1.06 : the complexity of a 32x64 window scan, for instance.

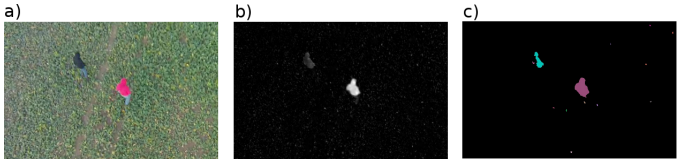

Fig. 6. a) input image, b) saliency map of the input image, c) ROIs extracted from the saliency map.

C. Analyse only the relevant locations

1) Principle: The exhaustive search of small objects is time-consuming. Reducing the search space will aid in decreasing the computation time. The analysis of the saliency map permits to extract regions of interest because people's saliency is high in uncluttered environments (Fig.6). However, this technique is not suitable when looking for closely situated people because the saliency becomes noisy and tricky to analyze. But in the last case there is no real need to reduce the search space because searching for closely situated people is quite fast.

The saliency map represents the saliency of a scene, with the concept being first introduced by Koch and Ullman [17]. The most salient locations extracted from this map are supposed to predict, quite well, the eye fixation locations. In uncluttered environments a small number of regions of interest can be deduced from these locations. We are particularly interested in the bottom-up saliency because it can be computed using only the pixel information [18]. But in return, the high-level context cannot be taken into account.

The search space reduction requires a fast and discriminative method. For this task we retained four different methods. Frintrop et al's method using integral images [19]. Their method is based on Itti et al's work [20]. Difference of gaussians are performed on several specific channels to mimic the visual receptive fields. The Katramados and Breckon's method [21]. They propose to divide two gaussian levels of two different gaussian pyramids to obtain the saliency map. The Achanta et al's method [22]. They substract the pixels of the gaussian blurred version of the image to the arithmetic mean pixel value of a maximum symmetric surround. Lu et al's method [23]: high-quality saliency map is deduced from image co-occurence histograms.

2) Pipeline: Reducing the search space using the saliency map requires a different approach, with a simple pipeline being proposed (Fig.7). The saliency map is first computed from the input image (Fig.7 2) and is then thresholded. Blobs of sufficient size are retained (Fig.7 3). Bounding boxes of all the retained blobs are computed (Fig.7 4). Detection windows are then generated for each blob and the centers of the windows are randomly chosen inside the bounding boxes (Fig.7 5). All the detection windows are treated separately and the results are fused with the mean-shift procedure (Fig.7 6). Some parameters have to be considered: the threshold value, the minimum number of pixels to retain a blob, the number of detection windows to generate and the scale range.

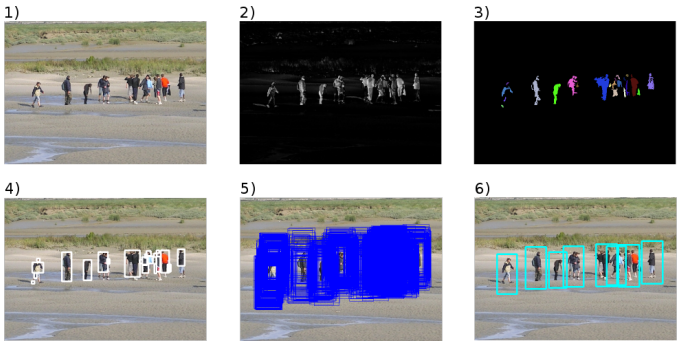

Fig. 7. The proposed pipeline for using the saliency. The saliency map is computed 2) from the input image 1), blobs are extracted 3), bounding boxes of the blobs are computed 4), detection windows are generated around the boxes 5) and results are fused 6).

The threshold value depends on the environment complexity. This value can be experimentally chosen. The minimum number of pixels for keeping a blob is a parameter used for filtering the noise. This value can be changed with respect to the distance to the people but most of the time a constant value permits sufficient filtering. The bigger the number of detection windows the better the detection, but the computation time will increase as well, and this must be taken into consideration when tuning this parameter.

IV. Testing

A. A more appropriate dataset

1) Methodology: At first a test was conducted to gradually compare the detector's response to elevation degrees greater than zero. A comparaison was made between the training efficiency of the two training datasets: the INRIA and the GMVRT-V1 datasets. The purpose of this test was to know from which elevation degree the INRIA trained detector started to produce worse results. Images of people, taken from elevations between 0 and 90degrees, were used.

Secondly, the global performance of the GMVRT-V1 trained detector was evaluated using the aerial test dataset. The purpose of this test was to confirm the suitability of such a multi-view training for human detection in real flight conditions.

2) Results: The average detection rate of the INRIA trained detector starts decreasing from 40degrees elevation and higher whilst the average detection rate of the GMVRT-V1 trained detector is not sensitive to the elevation (Fig.8). The GMVRT-V1 trained detector is therefore more robust to elevation. The

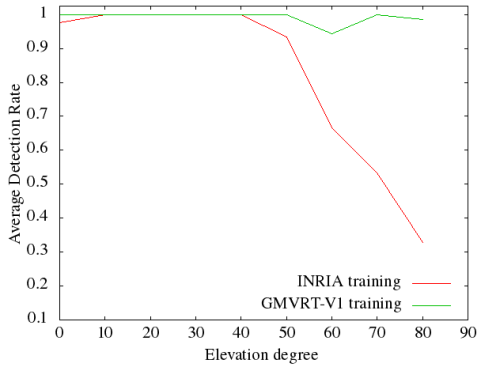

Fig. 8. Elevation robustness : the GMVRT-V1 training improves the robustness to the elevation compared to the INRIA training.

robustness has been tested on 180 cases. As far as we know the GMVRT-V1 training dataset is the first designed to manage elevation. Therefore, it is not possible to present comparaison results. However, one can easily see that the behaviour of the GMVRT-V1 trained detector, tested on the aerial dataset, is similar to the behaviour of pedestrian detectors in a pedestrian context [24] (Fig.9). The detection results are greatly improved with this training. The INRIA trained detector never obtains a better than 0.9 miss-rate when tested on the aerial test dataset.

It is believed that better results could be obtained by increasing the diversity of the positive training images and also by adding to the training dataset more various negative images of natural and uncluttered environments. It was also observed that the perspective effect tends to slightly rotate the shape of people according to the drone's altitude and this alters the detection results.

B. Smaller detection windows for detecting smaller objects

1) Methodology: The performance of three detectors were tested: the 32x64 HOG detector, the 48x96 HOG detector and the original 64x128 HOG detector. Each detector has been trained with INRIA training images resized to fit the window size. All the detectors have been bootstrapped once using full INRIA negative images resized according to the window size (three-quarter and one-half factor for the 48x96 and the 32x64 HOG detectors, respectively). The performance of the three detectors have been evaluated on a four times downsampled version of the INRIA test dataset. The image pyramids are configured as follows: 64 scales, a scale range from 0.6 to 4.65 for the 64x128 detector, a scale range from 0.45 to 3.2 for the 48x96 detector and a scale range from 0.3 to 2.13 for the 32x64 detector.

2) Results: The shapes of the 32x64 and the 48x96 curves are very similar to the shape of the 64x128 curve (Fig.10). According to Dollar's critera [24] the performance of the three detectors are close: the miss-rate at 1 FPPI (Falses Positives Per Images) is similar for the three detectors. We have similar detection performance using smaller windows whilst executing fewer pixel classifications.

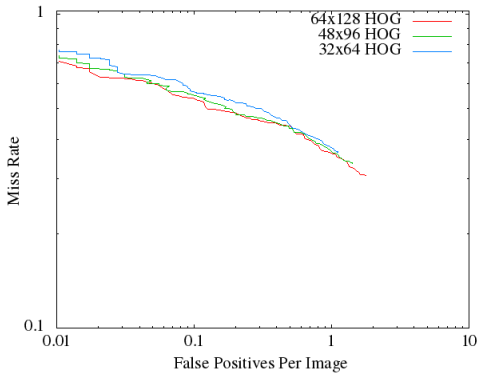

Fig. 10. ROC curves of three different detectors : the ROC curves of the 32x64 and the 48x96 detectors are similar to the ROC curve of the 64x128 detector.

C. Analyse only the relevant locations

1) Methodology: Firstly, by using the aerial test dataset, the global performance and computation time of three saliency algorithms were compared: the Achanta's algorithm [22], the Lu's algorithm [23] and the Katramados' algorithm [21]. It was chosen not to use the gradients' saliency within the Lu's algorithm in order to have similar behaviors. The total surface of all the bounding boxes containing the most salient regions is computed for each image. A reduction ratio is computed

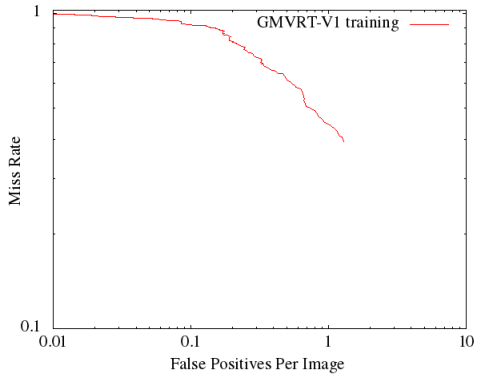

Fig. 9. ROC curve of the detector trained with the GMVRT-V1 dataset : the shape of the curve is comparable to the shape of curves of pedestrian detectors in pedestrian-view.

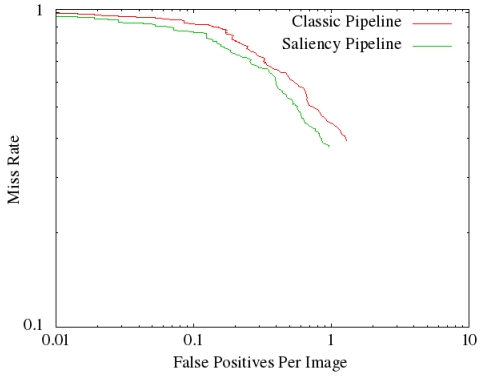

Fig. 11. ROC curves of two detectors using different pipelines: the ROC curve of the detector using the saliency pipeline is similar to the curve of the detector using the classic pipeline, results are slightly better.

TABLE I. SALIENCY REDUCTION EFFICIENCY COMPARAISON

Algorithm	Min (%)	Max (%)	Mean (%)	Sdev (%)	AvTime (s)
Achanta	0.0018	0.1857	0.0178	0.0227	1.34
Lu	0.0064	0.1625	0.0322	0.0339	4.02
Katramados	0.0027	0.6586	0.0401	0.0560	0.08

each time by dividing this surface by the image surface. The standard deviation and the mean of the reduction ratios are computed for each algorithm. The threshold and the minimum pixel number parameters have been tuned accordingly to the dataset to obtain the most of people's saliency and the least possible amount of noise. The chosen parameters are: a threshold of 0.18 and a minimum pixel number of 30 for Achanta's algorithm, a threshold of 0.20 and a minimum pixel number of 50 for Katramados's algorithm and a threshold of 0.25 and a minimum pixel number of 50 for Lu's algorithm.

Secondly, the global performance of two detectors, one using the saliency pipeline and the other the classic pipeline, were compared. It was chosen to use Achanta's algorithm in the saliency pipeline.

2) Results: Katramados' algorithm is the fastest (table I). It is also the most sensitive: the standard deviation (Sdev) is the biggest. And it is the one generating the most of noise : the mean is the biggest. Lu's algorithm is the slowest and it generates more noise than Achanta's. Achanta's algorithm has the best space reduction performance.

One does not necessarily choose the fastest saliency algorithm: the reduction ratio is a more important critera. A very fastly obtained but noisy saliency map leads to more blobs and the generation of unwanted windows, resulting in increased computation time.

The two curves of the Fig.11 have a similar shape. We observe that the curve of the detector using the saliency pipeline is below the other one. The performance of this detector are better. This can be explained because the detection windows are more closely treated. In the original pipeline, detection windows are treated every 8 pixels in X and Y (8 pixels is a cell-shift).

An average computation time of 18.879 sec per image is obtained with our implementation of the HOG detector and an average computation time of 2.286 sec per image using the saliency pipeline.

V. CONCLUSION

It has been shown that the multi-view training improves the detection results when the camera is directed to the ground with a pitching angle greater than 40degrees. Two solutions with which to reduce the computation time have been proposed: smaller detection windows and a new pipeline called saliency pipeline. The smaller detection windows proposed in this paper permit to perform fewer pixel classifications whilst having relatively similar detection performance. And the saliency pipeline speeds up the detection by a factor of nine in our case and it slightly improves the robustness. Implementation of the saliency algorithm on an FPGA could improve the speed factor to obtain real-time performance for detecting far off situated people.

REFERENCES

[1] J. Zhou and J. Hoang, "Real Time Robust Human Detection and Tracking System," in *Computer Vision and Pattern Recognition*, 2005.

[2] D. Toth and T. Aach, "Detection and recognition of moving objects using statistical motion detection and Fourier descriptors," in *International Conference on Image Analysis and Processing*, 2003.

[3] D. Gavrila and J. Giebel, "Shape-based pedestrian detection and tracking," *Intelligent Vehicle Symposium*, 2002.

[4] C. Papageoriou and T. Poggio, "A Trainable System for Object Detection," *International Journal of Computer Vision*, 2000.

[5] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Computer Vision and Pattern Recognition*, 2001.

[6] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Conference on Computer Vision and Pattern Recognition*, 2005.

[7] D. Lowe, "Object recognition from local scale-invariant features," in *International Conference on Computer Vision - Volume 2*, 1999.

[8] C. Wojek and B. Schiele, "A Performance Evaluation of Single and Multi-feature People Detection," *Pattern Recognition, 30th DAGM Symposium*, 2008.

[9] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral Channel Features," in *Proceedings of the British Machine Vision Conference*, 2009.

[10] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial Structures for Object Recognition," *International Journal of Computer Vision*, 2005.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models." *Transactions on pattern analysis and machine intelligence*, 2010.

[12] A. Gszczak, T. P. Breckon, and J. Han, "Real-time People and Vehicle Detection from UAV Imagery," in *Intelligent Robots and Computer Vision*, 2011.

[13] P. Rudol and P. Doherty, "Human Body Detection and Geolocalization for UAV Search and Rescue Missions Using Color and Thermal Imagery." in *Aerospace Conference*, 2008.

[14] V. Reilly, B. Solmaz, and M. Shah, "Geometric constraints for human detection in aerial imagery," in *European conference on Computer vision: Part VI*, 2010.

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

[16] P. Blondel, A. Potelle, C. Pégard, and R. Lozano, "How to improve the HOG detector in the UAV context," in *IFAC RED-UAS Workshop*, 2013.

[17] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *Transactions on pattern analysis and machine intelligence (TPAMI)*, 1998.

[18] A. Borji and L. Itti, "State-of-the-art in visual attention modeling." *Transactions on pattern analysis and machine intelligence (TPAMI)*, 2013.

[19] S. Frintrop, "VOCUS: A Visual Attention System for Object Detection and Goal-directed search," Ph.D. dissertation, 2010.

[20] L. Itti and C. Koch, "Computational modelling of visual attention." *Nature reviews. Neuroscience*, 2001.

[21] I. Katramados and T. Breckon, "Real-time visual saliency by division of gaussians," in *Internation Conference on Image Processing (ICIP)*, 2011.

[22] R. Achanta and S. Sabine, "Saliency Detection Using Maximum Symmetric Surround," in *Internation Conference on Image Processing (ICIP)*, 2010.

[23] S. Lu, C. Tan, and J.-H. Lim, "Robust and Efficient Saliency Modeling from Image Co-occurrence Histograms." *Transactions on pattern analysis and machine intelligence (TPAMI)*, 2013.

[24] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *Conference on Computer Vision and Pattern Recognition*, 2009.