Victim Detection from a low-altitude fixed-wing UAV: Experimental Results

Anurag Sai Vempati, Gabriel Agamennoni, Thomas Stastny, Roland Siegwart

No Institute Given

Abstract. This paper outlines a method to identify humans from a low-altitude fixed-wing plane relying on various Visual and Inertial sensors including an Infrared camera. The work draws inspiration from the need to detect victims trapped in a disaster situation in real-time and help the rescue units reach them. Such a work can also be easily employed for surveillance related applications. With recent advances in thermal imaging cameras, we now have the opportunity to explore more easier and reliable real-time applications. We start by pointing out various challenges that arise due to camera imperfections, viewpoint, altitude, synchronization etc. We provide a pipeline to efficiently fuse thermal and visual aerial imagery and get robust real-time detections. Confident detections are tracked across various frames and the real-time GPS locations of the victims is conveyed. Performance of our detection algorithm on various challenging datasets is provided.

1 Introduction

Search and Rescue is a widely researched field owing to it's numerous advantages in disaster scenarios. In cases like avalanches, a few minutes could make considerable difference in having better probability of suppressing the casualties. With increasing autonomy of the Unmanned Aerial Vehicles (UAV) and camera imaging technologies, it is now possible to scan large areas in a very short time and perform perception algorithms on-board at high rates with close to zero human intervention. Such technology also enables surveying in-accessible regions and hostile terrains making the task of dispatching rescue efforts considerably easy. In this paper we describe an algorithm that can efficiently detect trapped victims while autonomously scanning large areas using a UAV equipped with various sensors including visual and thermal cameras and an on-board computer to perform real-time computations. We will briefly outline individual components involved and provide results on a field experimental test.

Visual spectrum cameras have been used since a long time on UAVs but analyzing these images at high rates requires very robust algorithms to deal with various difficulties posed due to the size of objects of interest, motion blur, and viewpoint to name a few. Detecting humans from a UAV cruising at an altitude of 50-100 meters requires very high resolution cameras and the ability to quickly detect objects occupying few tens of pixels in area. On the other hand thermal cameras offer an advantage in such cases which makes it easier

to narrow down the search space to hotter objects. But thermal cameras have their own limitations like low Signal-to-Noise Ratio (SNR), white-black/hot-cold polarity changes, and halos that appear around very hot or cold objects [1]. We propose a technique that best utilises the pros of either cameras using sensor fusion technique.

Various works like... leverage on hotspot techniques to quickly narrow down the potential areas to further process. [2] uses fast-screening technique by modelling the background as an average of previous frames. For the foreground regions template called Contour Saliency Maps (CSM) are generated that preserves the edges that are both strong and significantly different from the background. The potential regions are found by correlating this template across the image. Such averaging techniques perform poorly in case of fast moving cameras like on UAVs. Many techniques rely on some kind of classifier to detect presence of a human in each of these potential areas. [2], [3], [4], [1] use some kind of variant of cascade of boosted classifiers introduced by Viola and Jones [5] - which basically involves a series of weak classifiers each better than the previous one. [6] show performance of various feature based classifiers trained on thermal data collected across wide variation in temperature, altitude and camera movement. HOG feature based classifier in conjunction with a particle filter tracker was found to perform the best.

Sensor fusion and Multi-modal image registration is a well explored field. But most of the works involving registration of Infrared and Visual camera images like [7], [8], [9] involves image processing techniques that rely on feature extraction, edge detection, segmentation etc. Techniques like such, though very robust, can be quite time-consuming for applications like ours. On the other hand, [4] uses camera intrinsics and ground planarity assumption to estimate relevant part of visual camera image for stationary victims. This method has it's own limitations due to additional criteria enforced on targets' positions and the environment being scanned. In our work we make use of camera extrinsics and use techniques from multi-view geometry to get one-to-one correspondence between Infrared and Visual images.

2 Victim Detection Pipeline

Detecting humans from an altitude of 50-100 meters with a camera of limited resolution poses a very challenging problem. Basic blocks of our pipeline are mentioned below:-

2.1 Datasets

2.2 Background Subtractor

Fig. 1 shows a human as seen in false-color rendering of the thermal camera image at an altitude of about 70 meters. At this scale, the humans occupy less than 50 pixels (<0.02%) in an image of 640 x 512 resolution. An exhaustive

search for such a tiny object of interest is very time consuming. So, we propose a Background Subtractor that returns regions of interest (ROI) and narrows down the search space considerably, thus enabling real-time detection. The foreground here is defined as a part of the image whose temperature differs quite significantly from it's surroundings. Since, humans are usually hotter (in winter) or colder (in summer) than the surroundings, we propose a technique that adaptively adjusts 2 threshold values (t_{low}, t_{high}) based on the surrounding pixel intensities. All the pixels with intensities less than t_{low} or greater than t_{high} are considered as foreground pixels.

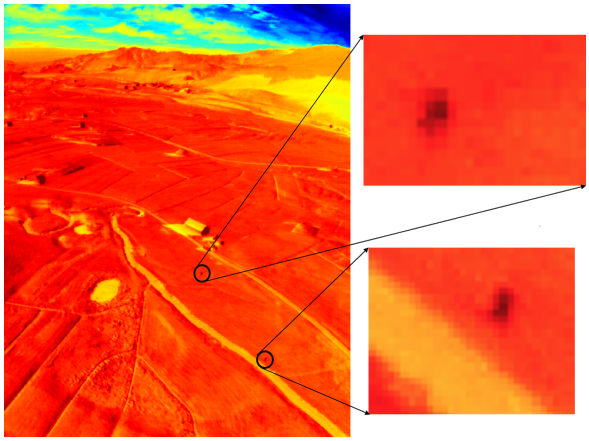
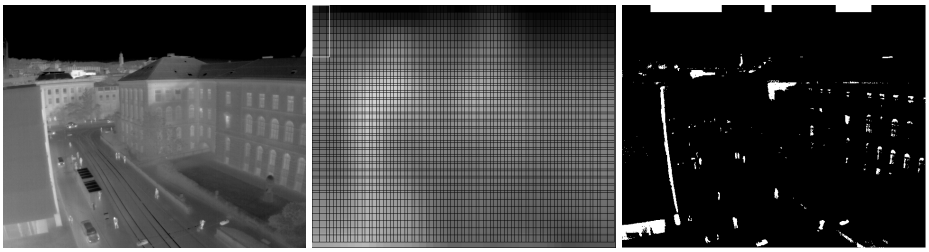


Fig. 1: A sample image recorded from the FLIR thermal camera at an approximate altitude of 70 meters. Two humans are pointed out.

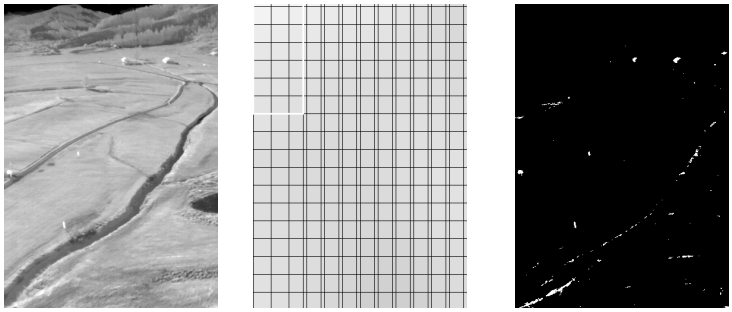
We employ a very simple sliding window based approach to estimate adaptive thresholds. The image is divided into various overlapping blocks and the adaptive thresholds t^b_{low}, t^b_{high} for each block b is evaluated as the higher and lower quantiles of the Gaussian models fitted to the pixel intensities of the block. Now, the threshold values at each pixel location (i, j) are chosen as weighted average of all the block thresholds t^b_{low}, t^b_{high} if the pixel belongs to block b. The weights are chosen inversely proportional to the pixel's distance from the block's center. A segmentation map (segmap) is generated by thresholding the Infrared image at each pixel location.

Fig.?? shows 2 sample images collected at different altitudes, temperature and times of the day and the corresponding segmaps. A blob detection algorithm [10] is used to find blobs of desired size, thus providing ROIs to search for humans. For the UAV scenario, we further narrow down the search space by considering only the blobs whose area lies in certain range that is calculated at every time-step based on the camera specifications, mounting, UAV's IMU pose and GPS position. This approach is computationally much cheaper and is not affected by

the fast moving camera. It provides much better results at real-time than our previous implementation using Vibe [11] which was designed for visual images.



(a) IR image (b) 8-bit threshold image (c) Segmentation map



(a) IR image (b) 8-bit threshold image (c) Segmentation map

Fig. 2: Sequence 1 is captured in the night time at an altitude of 20 meters above the ground with some crowd. Sequence 2 was captured from AS[??] on a cold winter morning at an altitude of 50 meters above the ground. (a) shows the Infrared image, (b) is the 8-bit image of pixel-wise threshold $t_{high}(i, j)$ (grid of blocks is overlaid and a sample block is marked in white on the top-left) and (c) is the resulting segmentation map depicting foreground pixels.

2.3 Human Classifier

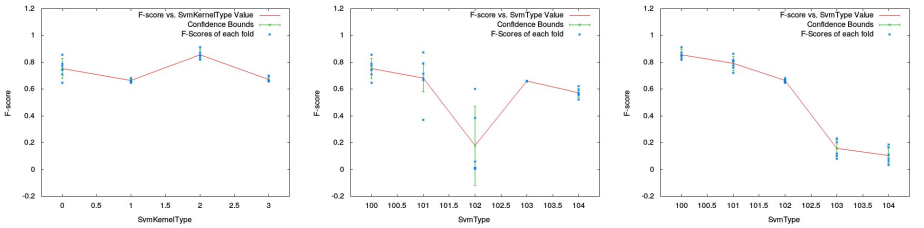
The ROIs obtained from Background Subtraction are exhaustively searched for presence of a human. For this, we use a HOG feature based learning classifier (which was found to be optimal [6]) to classify the extracted patch into human or non-human category. The training data is obtained by manually annotating sequences from different datasets using vBBToolbox [12]. The distribution of training data across the datasets is shown in Table??. An SVM classifier is trained using these image patches. Type of SVM optimization problem, kernel type and other parameters are optimized using K-fold cross validation. Fig. 3 shows performance of few chosen configurations. C_SVC optimization problem

with Linear kernel type and RBF kernel type are found to be 2 top performing classifier configurations.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes . . .**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Headings.** Text follows . . .	10 point, bold
4th-level heading	*Remark.* Text follows . . .	10 point, italic

Table 1: Distribution of training image patches among different datasets

HOG features are extracted on the image patches at multiple scales and the descriptor generated is passed on to the SVM classifier to detect the presence of humans. Fig.?? shows PR curves for the performance of classifier on CLA data.



(a) Choosing kernel type for C_SVC problem (b) Choosing optimization problem for linear SVM kernel (c) Choosing optimization problem for RBF SVM kernel

Fig. 3: Performance of SVM classifier with different configurations of optimization problem and kernel type. It's found to be optimal to choose (a) RBF kernel type for C_SVC problem, (b) C_SVC problem for Linear kernel type and (c) C_SVC problem for RBF kernel type.

¡one-class SVM¿

2.4 Infrared-Grayscale Image Fusion

Objects in Infrared image rarely have discerning features since the temperature tends to be uniform across the objects. So, a classifier would be not so reliable in classifying such patches since it only has the shape information of the object in a very small patch. Since in all our flights using the sensorpod we collected both Infrared and Grayscale images, a fusion technique is proposed to get a one to one relationship between pixel locations of Infrared and Grayscale images. This will

help us in extracting both Infrared and Grayscale image patches corresponding to the ROIs obtained from Background Subtraction which can help the classifier in producing more reliable results.

We use Kalibr [13] to calibrate the Infrared and Grayscale camera intrinsics and extrinsics for the sensorpod mounting. We use radial-tangential model to estimate camera distortion parameters. An april tag pattern is used to estimate Grayscale camera intrinsics. A thick matte paper with checkerboard pattern under illumination of a bright light is used to calibrate Infrared camera intrinsics. We then undistort the images and choose a new camera matrix for both the cameras with common focal parameters. An optimal camera matrix is evaluated that ensures all the original pixels are present in the final image after undistorting the images. Rotation and Transaltion matrices between the cameras is obtained using camera-camera extrinsics evaluated with Kalibr. Now the Infrared and Grayscale camera setup can be treated as a Stereo pair and standard image rectification techniques can be used to evaluate the necessary projection matrices. At this point, all the epipolar lines are parallel to image edge (horizontal if horizontal rectification was employed, vertical otherwise). For our particular setup, vertical stereo was chosen since the camera centers were aligned vertically on the sensorpod. Now that the disparities of the image are in one axis, it's fast and easy to find pixel correspondences. After a manual displacement that is set for all the columns, individual column displacements are evaluated using correspondences between fast features calculated for both the images.

Fig. 5 shows results from the fusion technique. In Rothethurm-1 sequence IR and Grayscale cameras were both facing in the direction of flight and are 25° nadir configuration with a translation of 3.5cm between the camera centers. In Sea-1 sequence, IR camera was in 25° nadir configuration while Grayscale camera was in 50° nadir configuration with a translation of 3.5cm between the camera centers. The results show reasonable overlap accuracy. Since, we only require a bounding box containing the same object in both Infrared and Grayscale image, this level of accuracy suffices.

Fig.?? shows improvement to the PR curves after Fusion on Roth-1 data.

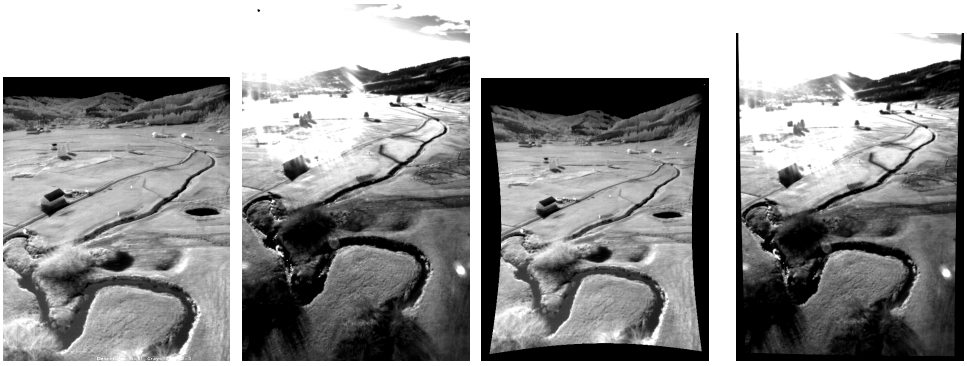
3 Experimental Results

3.1 Structural Inspection Planner

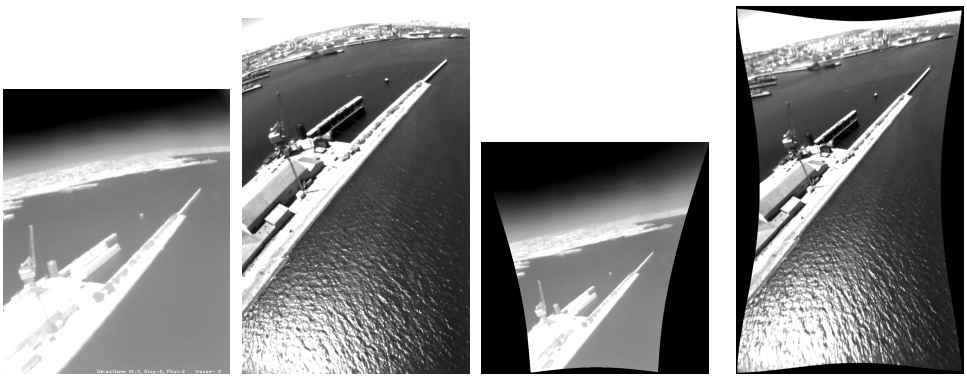
3.2 Estimating victim GPS

Fig. 6(a) shows a schematic of all the involved sensors, the corresponding axes and the intrinsics/extrinsics involved. For estimating the GPS positions of detections, we assume the ground as a plane. Using the calibration data obtained earlier, it can be easily shown the latitude/longitude are:

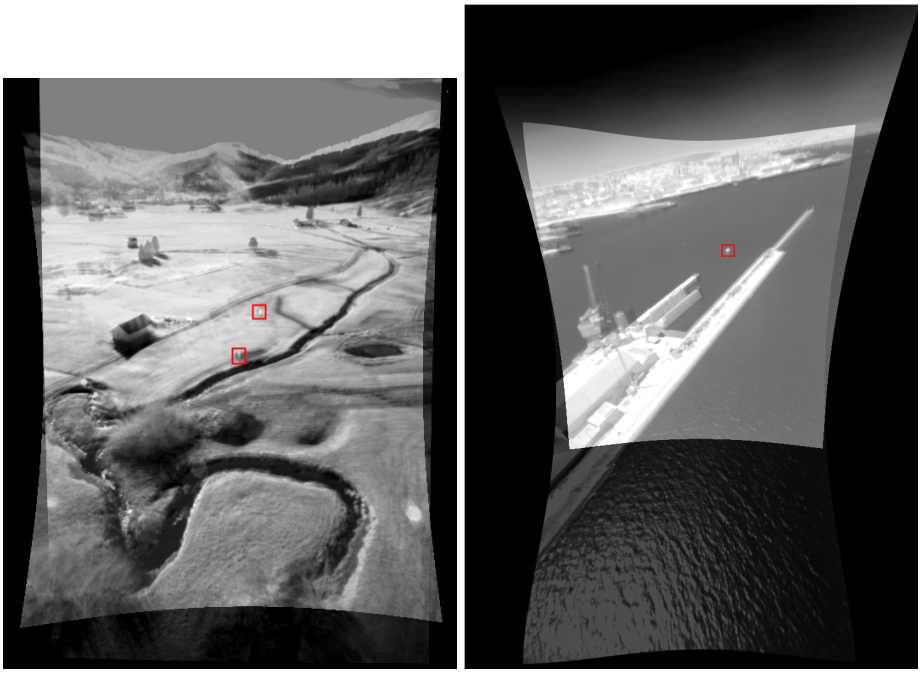
Fig. 6(b) shows the detected victim GPS positions, actual groundtruth victim location and the path followed by the UAV. The rest of the detections include unregistered humans in the explored area and some False Positives. By enforcing a criterion - similar to [4] - that requires a particular GPS position to be



(1.a) IR image (1.b) Grayscale Image (1.c) Rectified IR (1.d) Rectified Grayscale

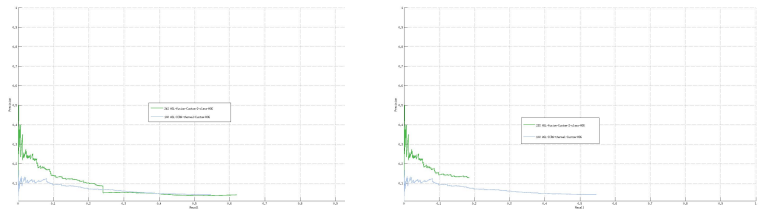


(2.a) IR image (2.b) Grayscale Image (2.c) Rectified IR (2.d) Rectified Grayscale



(1.e) Fused IR and Grayscale Images (2.e) Fused IR and Grayscale Images

Fig. 4: The top 2 rows show 2 sequences (Rothenthurm-1, Sea-1) of original images obtained from IR and Grayscale cameras and the resulting images post rectification. Last row shows fusion of 2 images to visualize the quality of overlap. Red boxes in seq-1 highlights 2 humans and a boat in seq-2. As can be seen, the correspondence between IR and Grayscale images is reasonably good.



(a) PR curves when all the detections are retained

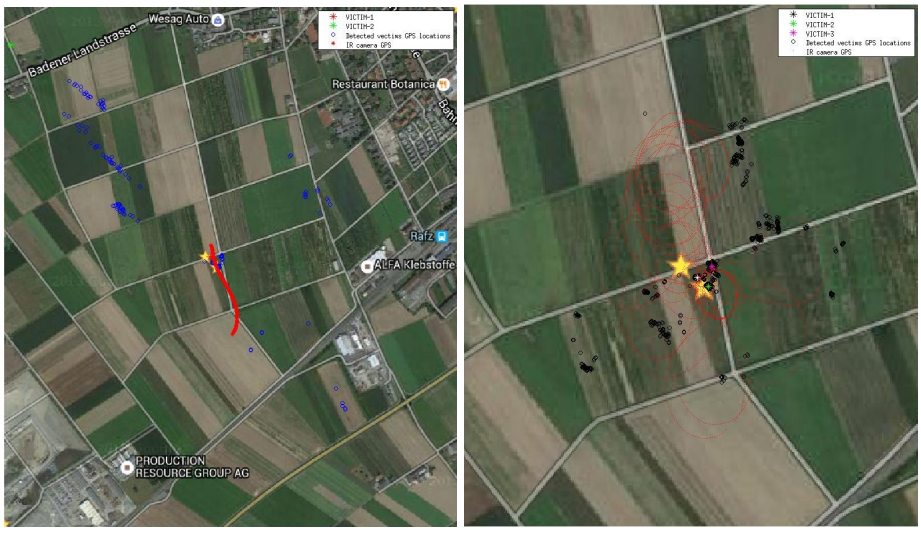
(b) PR curves when only 2 best detections per frame are retained

Fig. 5: The top 2 rows show 2 sequences (Rothenthurm-1, Sea-1) of original images obtained from IR and Grayscale cameras and the resulting images post rectification. Last row shows fusion of 2 images to visualize the quality of overlap. Red boxes in seq-1 highlights 2 humans and a boat in seq-2. As can be seen, the correspondence between IR and Grayscale images is reasonably good.

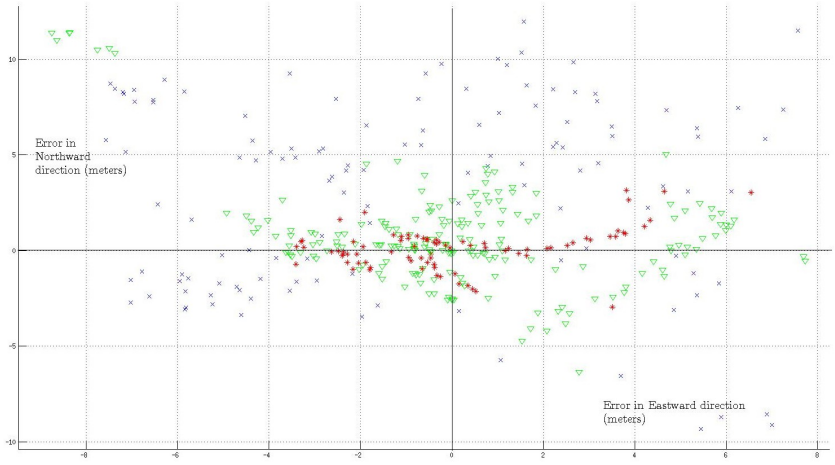
detected consistently within the duration of the time that particluar spot is observed, most of the Flase Positives are eliminated and the set of true detections are registered to known groundtruths. False Positives are further eliminated by boosting the confidence of detections that are closer to the GPS location of the point where principal axis of camera hits the ground plane. Fig. 6(c) is a plot of GPS position estimate errors for all the 3 registered victims. The GPS estimates are converted to UTM coordinates to represent the error in meters. The errors obtained are a cumulative of errors in groundtruth GPS measuring device, calibration imperfections, GPS to/from UTM conversions, images' resolution, planarity assumptions, intertial sensor imperfections (GPS, IMU). So, it's quite uncertain how much of the error can be attributed to the imperfections in our algorithm. The only assumption made in our approach is of ground planarity. Even this can be easily overcome by using geographical elevation data of the scanned region.

References

1. Wang, W., Zhang, J., Shen, C.: Improved human detection and classification in thermal images. In: Image Processing (ICIP), 2010 17th IEEE International Conference on, IEEE (2010) 2313–2316
2. Davis, J.W., Keck, M.A.: A two-stage template approach to person detection in thermal imagery. In: null, IEEE (2005) 364–369
3. Treptow, A., Cielniak, G., Duckett, T.: Active people recognition using thermal and grey images on a mobile security robot. In: Intelligent Robots and Systems, 2005. (IROS 2005). 2005 IEEE/RSJ International Conference on. (2005) 2103–2108
4. Rudol, P., Doherty, P.: Human body detection and geolocalization for uav search and rescue missions using color and thermal imagery. In: Aerospace Conference, 2008 IEEE, IEEE (2008) 1–8



(a) Victim GPS for a small segment (b) Victim GPS



(c) Victim UTM position errors

Fig. 6: asdasdsadsada.

5. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Volume 1., IEEE (2001) I–511
6. Portmann, J., Lynen, S., Chli, M., Siegwart, R.: People detection and tracking from aerial thermal views. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on, IEEE (2014) 1794–1800
7. Toet, A., van Ruyven, L.J., Valeton, J.M.: Merging thermal and visual images by a contrast pyramid. Optical Engineering **28** (1989) 287789–287789–
8. Heo, J., Kong, S., Abidi, B., Abidi, M.: Fusion of visual and thermal signatures with eyeglass removal for robust face recognition. In: Computer Vision and Pattern Recognition Workshop, 2004. CVPRW '04. Conference on. (2004) 122–122
9. Istenic, R., Heric, D., Ribaric, S., Zazula, D.: Thermal and visual image registration in hough parameter space. In: Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on. (2007) 106–109
10. nán, C.C.L.: cvBlob. (http://cvblob.googlecode.com)
11. Barnich, O., Van Droogenbroeck, M.: Vibe: A universal background subtraction algorithm for video sequences. Image Processing, IEEE Transactions on **20** (2011) 1709–1724
12. Dollár, P.: Piotr's Computer Vision Matlab Toolbox (PMT). (http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html)
13. Furgale, P., Rehder, J., Siegwart, R.: Unified temporal and spatial calibration for multi-sensor systems. In: Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on, IEEE (2013) 1280–1286