Silhouette Coverage Analysis for Multi-modal Video Surveillance

Steven Verstockt, Chris Poppe, Pieterjan De Potter, Charles Hollemeersch, Sofie Van Hoecke, Peter Lambert, Rik Van de Walle

Proceedings of the 29th Progress in Electromagnetics Research Symposium (PIERS), pp. 1279-1283, 2011.

Silhouette Coverage Analysis for Multi-modal Video Surveillance

S. Verstockt[1, 2], **C. Poppe**[1], **P. De Potter**[1], **C. Hollemeersch**[1],
S. Van Hoecke[2], **P. Lambert**[1], **and R. Van de Walle**[1]

[1]ELIS Department Multimedia Lab, Ghent University, IBBT, Belgium
[2]ELIT Lab, University College West Flanders, Ghent University Association, Belgium

Abstract— In order to improve the accuracy in video-based object detection, the proposed multi-modal video surveillance system takes advantage of the different kinds of information represented by visual, thermal and/or depth imaging sensors. The multi-modal object detector of the system can be split up in two consecutive parts: the registration and the coverage analysis.

The multi-modal image registration is performed using a three step silhouette-mapping algorithm which detects the rotation, scale and translation between moving objects in the visual, (thermal) infrared and/or depth images. First, moving object silhouettes are extracted to separate the calibration objects, i.e., the foreground, from the static background. Key components are dynamic background subtraction, foreground enhancement and automatic thresholding. Then, 1D contour vectors are generated from the resulting multi-modal silhouettes using silhouette boundary extraction, cartesian to polar transform and radial vector analysis. Next, to retrieve the rotation angle and the scale factor between the multi-sensor image, these contours are mapped on each other using circular cross correlation and contour scaling. Finally, the translation between the images is calculated using maximization of binary correlation.

The silhouette coverage analysis also starts with moving object silhouette extraction. Then, it uses the registration information, i.e., rotation angle, scale factor and translation vector, to map the thermal, depth and visual silhouette images on each other. Finally, the coverage of the resulting multi-modal silhouette map is computed and is analyzed over time to reduce false alarms and to improve object detection.

Prior experiments on real-world multi-sensor video sequences indicate that automated multi-modal video surveillance is promising. This paper shows that merging information from multi-modal video further increases the detection results.

1. INTRODUCTION

The growing demand for security has given raise to the increased use of video surveillance systems in recent years. Surveillance cameras are rapidly appearing in all sort of places and a huge number of visual object detection algorithms, which automatically process these camera images, have been proposed in literature. However, due to the variability of shape, motion, colors, and patterns of moving objects, and also due to the dynamic character of the background, many of these visual object detectors are still vulnerable to false and missed detections. To avoid the disadvantages of using visual sensors alone, we believe the use of other types of imagery, e.g., thermal infrared (IR) and Time-of-Flight (ToF) depth images, can be of added value. The combination of this types of imagery yields information about the scene that is rich in color, motion, depth and/or thermal detail. Once such information is registered, i.e., aligned with each other, it can be used to improve detection performance and activity analysis in the scene. Since each type of sensor has its own type of detection limitations, misdetections in one sensor can be corrected by the other sensors. As such, the combination of multi-sensor information is considered to be a win-win.

In order to combine multi-modal images, it is required that the corresponding objects in the scene are aligned, or registered. The goal of registration is to establish geometric correspondence between the images so that they may be transformed, compared, and analyzed in a common reference frame. Usual features used for multi-sensor registration are edges, corners, and contours [1]. Since contours representing the region boundaries are preserved in most cases, object silhouettes form the most reliable correspondence between objects in color, thermal and/or depth image pairs [2]. For this reason, they are also used in our multi-modal video surveillance system.

The remainder of this paper is organized as follows. Section 2 gives a global description of the silhouette-based registration of multi-modal images, which is based on moving object silhouette extraction, contour vector generation, contour mapping and binary correlation. As an example, the registration of visual and long-wave infrared (LWIR) images is shown. Subsequently, Section 3 discusses the silhouette coverage analysis, i.e., the multi-modal merging of the detection results

from the visual, thermal and/or depth image sensors. By two use cases, i.e., a shadow removal and a smoke detection experiment, we show how the coverage analysis of multi-modal images can be used to obtain better object detection results than either sensor alone. Next, in Section 4, we provide details of the experimental setup. Finally, Section 5 ends this paper with the conclusions.

2. SILHOUETTE-BASED REGISTRATION OF MULTI-MODAL IMAGES

The multi-modal image registration (Fig. 1) starts with a moving object silhouette extraction [2] to separate the calibration objects, i.e., the moving foreground, from the static background. Key components are the dynamic background subtraction, automatic thresholding and (iterative) morphological filtering. The dynamic background subtraction [3] extracts the moving foreground (FG) out of the visual and thermal video frames using a visual background estimation, which is updated dynamically. By subtracting the frames with everything in the scene that remains constant over time, i.e., the background, only the moving part of those images remains. After this background subtraction, the resulting foreground images are thresholded automatically using automatic gamma correction, (adaptive) k-means clustering and morphological filtering with growing structuring elements, which grow iteratively until the resulting silhouette is suitable for multi-modal silhouette matching. The combination of all these steps achieves favorable results, as is shown by the visual and the LWIR silhouette extraction in our experiments (Fig. 4). Similar results can be expected for ToF depth silhouette extraction.

After the silhouettes are extracted, registration of both images is performed using a three step registration algorithm. Like in [4], the registration algorithm assumes that the geometric transformation between the multi-sensor images is a rigid transformation, which can be decomposed into a 2D rotation, scaling and translation. To estimate each of these three geometric parameters, the contours and the correlation of the visual and thermal silhouettes are analyzed. First, the rotation is computed using silhouette contour extraction and circular cross correlation [5], which analyzes the translation of the 1-D contour centroid distance (CCD) of both silhouettes. As such, the 2D silhouette matching problem is converted to a one-dimensional signal matching problem. After rotating, the scale factor between both views is estimated by analyzing the ratio of the thermal

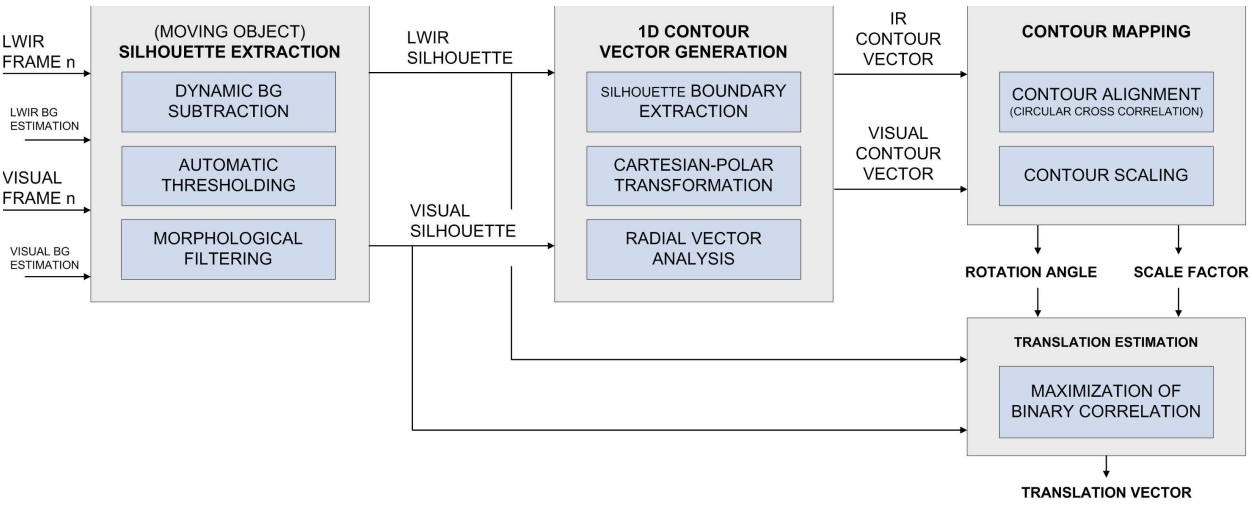


Figure 1: Silhouette-based image registration of thermal and visual images.

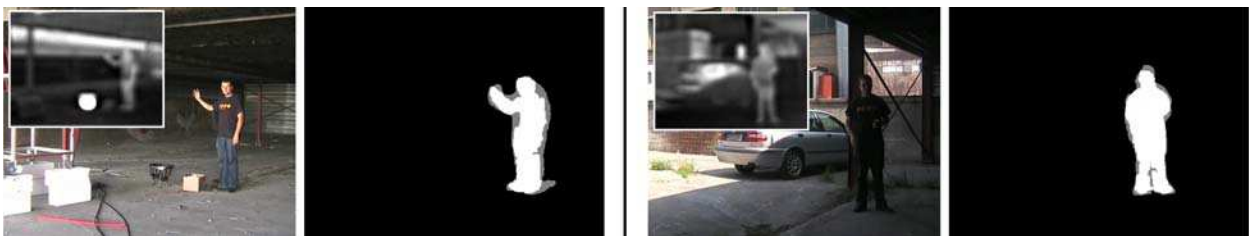


Figure 2: Experimental results of LWIR-visual registration.

and visual aligned CCDs. Since the thermal-visual CCD ratios are not constant and show some kind of disorder, the median ratio is chosen as an adequate scale factor. Finally, the translation vector is estimated using the binary correlation technique proposed by Chen et al. [2], which is based on template matching in the frequency domain. As the registration result in (Fig. 2) show, the proposed registration algorithm is able to coarsely map visual and thermal object silhouettes.

3. SILHOUETTE COVERAGE ANALYSIS

The silhouette coverage analysis (Fig. 3) also starts with the moving object silhouette extraction, which was already discussed in the previous section. Then, it uses the registration information, i.e., rotation angle, scale factor and translation vector, to map the thermal and visual silhouette images on each other. As soon as this mapping is finished, the combined LWIR-visual silhouette map is analyzed over time using a temporal coverage analysis algorithm. Depending the video surveillance application for which the multi-modal analysis is used, this silhouette coverag analysis (SCA) can be performed in different ways. In the following subsections, two exemplary use cases of how the SCA can be used are given. In the first use case, the SCA is used for shadow removal in visual images. In the second use case, the SCA is used as a first warning method for smoke detection.

3.1. Use Case 1: Shadow Removal

Shadows are a main drawback for all visual surveillance applications and affect the accuracy of the system performance. Since shadows do not occur in thermal or ToF depth images, both types of imagery can be used to discard them in visual images. This is also shown by the first experiment in (Fig. 4(a)). In this experiment, the multi-modal SCA is used to count the number of people in a room. Due to their shadows, the visual silhouettes of both persons overlap in the visual images. Without the LWIR-visual SCA, a visual people counter could miscount the number of people as 1. By using the LWIR-visual SCA we can correct this mistake. As can be seen in (Fig. 4(a)), the registered visual and thermal silhouettes do not overlap in the shadow regions, i.e., the gray

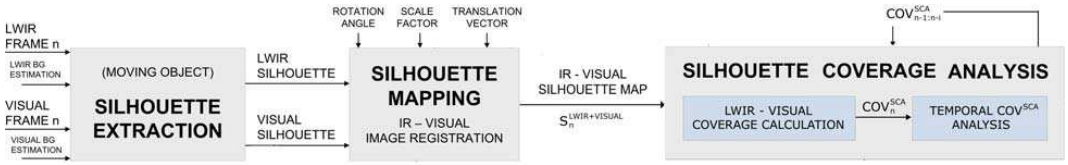
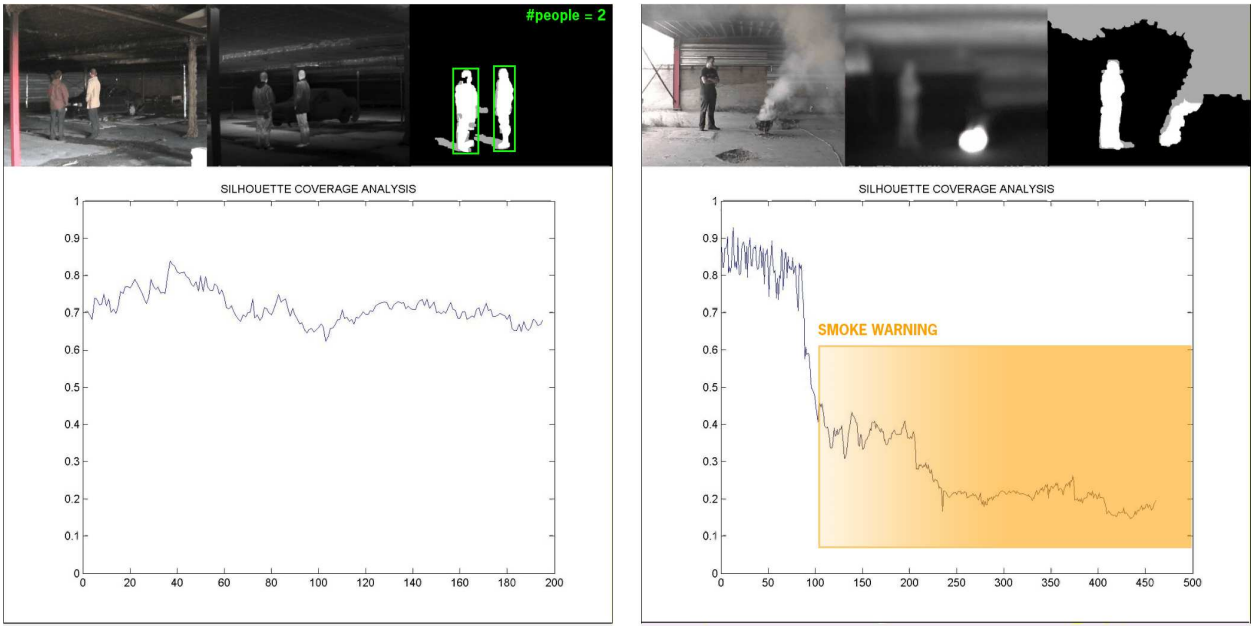


Figure 3: Silhouette coverage analysis.



(a) (b)

Figure 4: Experimental results of silhouette coverage analysis for (a) shadow removal and (b) smoke detection.

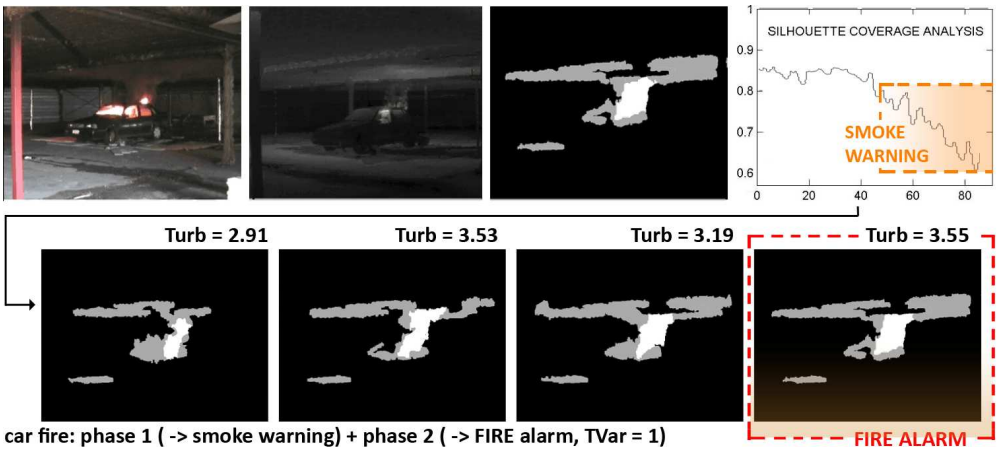


Figure 5: "Car park fire [8]" test results of SCA-based smoke detection.

regions. As such, by only counting the regions which occur in both thermal and visual images, i.e., the white regions, and by analyzing if this regions are stable over time, the SCA results in a more robust and efficient people counter. Similar results are expected with visual-ToF depth SCA analysis. The bounding boxes, shown in the figure, were created by calculating the smallest enclosing rectangle (whose sides are parallel to the x and y axes) around the common, i.e., white, visual-LWIR regions.

3.2. Use Case 2: Smoke Detection

Although smoke is almost transparent in LWIR images, we can make use of its absence to detect it. Since ordinary moving objects, such as people, cars, etc., produce similar silhouettes in background-subtracted visual and thermal IR images, the coverage between these images is quasi constant. This can also be seen in the coverage graph of experiment 1. The coverage for the moving people stays quasi constant over all the frames. Smoke, contrarely, will only be detected in the visual images, and as such the coverage will start to decrease (Fig. 4(b)). This decrease can be detected using a sequence/scene independent technique based on slope analysis of the linear fit, i.e., trend line, over the most recent silhouette coverage values. If the slope of this trend line is negative and decreases continuously, smoke warning is given. Due to its dynamic character, the visual silhouettes of a smoke region will also show a high degree of turbulence [6]. By focusing on both the visible-invisble character of smoke and its visual disorder, a multi-sensor detector can detect smoke very accurately.

Compared to the results of any individual detector in [7], the 2-phase multi-sensor smoke detector is able to detect the smoke more accurate, i.e., with less missdetections and false alarms. This is also illustrated by the test results of a car park fire [8] in (Fig. 5). Due to the low-cost of the silhouette coverage analysis and the visual disorder analysis, which is only performed if smoke warning is given, the algorithm is also less compuational expensive as many of the individual detectors.

4. EXPERIMENTAL SETUP

The multi-modal sequences were acquired by a Xenics Gobi-384 LWIR camera and a CANON MD110 camera, which works in the 8–14 µm spectral range and the visible spectrum respectively. The Gobi thermal imager has a resolution of 384 × 288 pixels, and a frame rate of 28–30 fps. The CANON its resolution is 576 × 720 and its framerate is 25 fps. In order to cope with the different frame rates and resolutions, and also whith the differences in the the field of view of the cameras, the multi-modal frames are spatio-temporal registered using temporal frame alignment and the silhouette-based registration proposed in this paper.

5. CONCLUSIONS

Multi-modal video surveillance takes advantage of the different kinds of information represented by thermal, visual and/or depth images in order to accurately detect moving objects. By fusing the different modalities and using the strengths of each medium, object detection can be done more accurate and with less false detections, as is shown by two use cases in this paper. Merging information from multiple types of image sensors has, as such, proven to be a win-win.

To detect the presence of objects, the detector analyzes the silhouette coverage of moving objects in multi-modal registered images. In order to register the multi-sensor images, the proposed algorithm analyses the contours and the correlation of visual and thermal FG silhouettes. First, the rotation is computed using silhouette contour extraction and circular cross correlation. Next, contour scaling is used to estimate the thermal-visual scale factor. Finally, the translation vector is estimated by maximization of binary correlation.

The geometric parameters found during this registration phase are further used by the detector to coarsely map the silhouette images and coverage between them is calculated. Depending the video surveillance application for which the multi-modal analysis is used, this coverage can then be further used to improve the detection results, as is shown by the people counter and the smoke detection experiment.

Future work will mainly focus on the improvement of the registration results. Currently, only the binary silhouettes of the calibration objects are used to do the registration. We expect that a first improvement can be made by also incorporating their gray-scale information, especially in the translation estimation. As the contour mapping is based on the boundary correspondences it is not expected that grayscale information will lead to better results in the rotation and scale estimation. However, further testing is necessary to confirm this. Also the use other types of class clustering classifiers will be further investigated.

ACKNOWLEDGMENT

The research activities as described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), University College West Flanders, Warrington Fire Ghent, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders, the Belgian Federal Science Policy Office (BFSPO), and the European Union.

REFERENCES

1. Zitova, B. and J. Flusser,, "Image registration methods: A survey," *Image and Vision Computingn*, Vol. 21, 977–1000, 2003.
2. Chen, H.-M., S. Lee, R. M. Rao, M.-A. Slamani, and P. K. Varshney, "Imaging for concealed weapon detection," *IEEE Signal Processing Magazine*, 52–61, March 2005.
3. Toreyin, B. U., Y. Dedeoglu, U. Gdkbay, and A. E. Cetin, "Computer vision based method for real-time fire and flame detection," *Pattern Recognition Letters*, Vol. 27, 49–58, 2006.
4. Han, J. and B. Bhanu, "Fusion of color and infrared video for moving human detection," *Pattern Recognition*, Vol. 40, 1771–1784, 2007.
5. Hamici, Z., "Real-time pattern recognition using circular cross-correlation: A robot vision system," *International journal of Robotics and Automation*, Vol. 21, pp 174–183, 2006.
6. Verstockt, S., A. Vanoosthuyse, S. van Hoecke, P. Lambert, and R. van de Walle, "Multi-sensor fire detection by fusing visual and non-visual flame features," *International Conference on Image and Signal Processing (ICISP)*, 333–341, June 2010.
7. Verstockt, S., B. Merci, B. Sette, P. Lambert, and R. van de Walle, "State of the art in vision-based fire and smoke detection," *14th International Conference on Automatic Fire Detection*, Vol. 2, 285–292, September 2009.
8. Merci, B., "Fire safety and explosion safety in car parks," http://www.carparkfiresafety.be/.