FUSION OF INFRARED AND VISIBLE SPECTRUM VIDEO FOR INDOOR SURVEILLANCE

C. Ó Conaire, E. Cooke, N. O'Connor, N. Murphy, A. Smeaton

Centre for Digital Video Processing, Adaptive Information Cluster, Dublin City University Ireland
{oconaire, oconnorn}@eeng.dcu.ie

ABSTRACT

In this paper, we describe an approach to video object segmentation using combined analysis of visible spectrum and far infrared imaged data captured using a novel camera rig. Combined infrared-visible spectrum analysis can produce higher quality object segmentation results than those possible when only one modality is considered, as well as being very robust to lighting changes that severely affect traditional surveillance systems. The presented approach uses adaptive filtering and thresholding of infrared data coupled with background modeling and change detection in colour video sequences. To illustrate the effectiveness and application of the approach, a prototypical surveillance system is described that detects when a person has entered a restricted area, even in total darkness, using combined analysis of infrared and visible spectrum video of an indoor scene.

1. INTRODUCTION

Visual surveillance is currently a very active research area and incorporates many computer vision techniques such as image and video analysis, object recognition and data fusion, as well as machine learning techniques. Hu et al [1] conduct an extensive survey on the state of the art in visual surveillance. They concentrate on the surveillance of people or vehicles, noting that they are typical of surveillance applications in general. Interestingly, an important trait that people and vehicles share is that their temperature is typically different to the background, thus by using infrared imaging, people and vehicles can be extracted and tracked more efficiently. As the technology develops, thermal infrared imaging devices are becoming more common and have been used in a variety of research areas. The use of infrared in pedestrian detection to reduce night-time accidents is investigated in [2] and [3]. A very comprehensive overview of image processing techniques and their application to infrared imagery is described in [4].

In this paper, we demonstrate that combined infrared-visible spectrum analysis can produce higher quality object segmentation results than those possible when only one modality is considered, as well as being very robust to lighting changes that severely affect traditional surveillance systems. In section 2, we describe the novel camera rig that allows the simultaneous capture of infrared and visible spectrum video. Section 3 describes our algorithm which uses adaptive filtering and thresholding of infrared data coupled with background modeling and change detection in colour video sequences to segment video objects. Section 4 provides example segmentation results, including segmentation in total darkness and some results from our prototypical surveillance system, which detects entry into a restricted area.

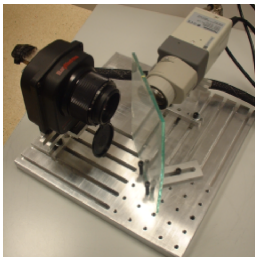

Fig. 1. Visible/Infrared Camera Rig

2. HARDWARE CONFIGURATION

Figure 1 shows the configuration of the visible and thermal cameras. A pane of standard window-glass was positioned between them to act as a beam-splitter. There are other more suitable types of material but we wanted to demonstrate that standard glass would be effective enough. The glass absorbs a significant portion of the infrared radiation and reduces the infrared image contrast as a result, but it solves the image registration problem of aligning the visible and infrared images, creating a four-band image. An automatic alignment technique that can be used for images of very different modalities (such as thermal and visible images) is proposed in [5]. They make an assumption that the scene is planar but this assumption does not hold in the indoor environments that this paper is concerned with. We use a Raytheon ControlIR 2000B thermal imaging video camera that is sensitive to wavelengths of 7μm-14μm, along with a Panasonic WV-CP470 video camera. The two cameras are synchronised (*gen-locked*) to ensure that they capture frames simultaneously. The analogue video output is captured and digitised by a Falcon Quattro multi-channel frame-grabber. The video frames from the visible and infrared bands were aligned using a planar homography [6].

3. ALGORITHMIC DETAILS

Figure 2 shows a simplified diagram of our system architecture, which is described in detail in this section.

3.1. Visible Background Modelling

To extract regions of interest in the visible spectrum, background modelling is used to detect pixels that are new or unusual and are thus classified as foreground pixels. The background-modelling algorithm used in this paper is based on [7], which is an improvement of the method described in [8]. The improved algorithm de-

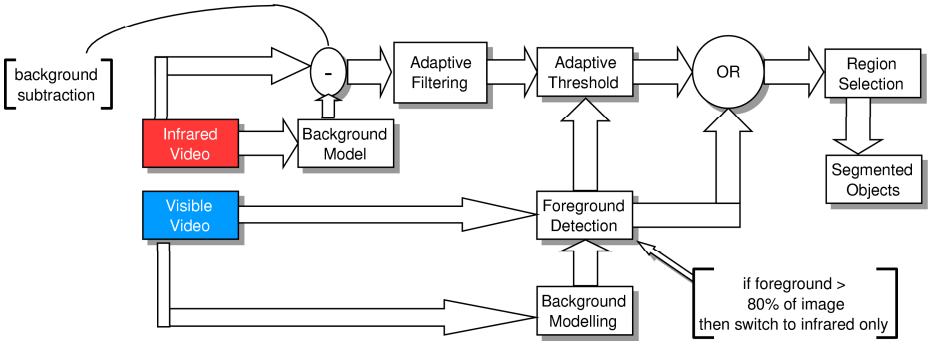

Fig. 2. System Diagram

creases the background model initialisation time substantially by estimating the Gaussian model parameters using the expected sufficient statistics equations and then switching to the L-recent window update equations when the first L frames have been processed. In the background model, each pixel is modelled by a mixture of K Gaussian distributions with the assumption that the red, green and blue components are independent and have the same variance. This assumption is designed to decrease the computational complexity. Based on the weights of the Gaussian distributions, a subset of them is chosen to represent the pixel's background distribution. A pixel from a new frame is compared to each of the Gaussians that make up the background model. If it is not within 2.5 standard deviations of any of them, it is classified as a foreground pixel. Foreground pixels may be reclassified as shadows if the colour distortion is small and the brightness has decreased slightly. We modified the algorithm so that the variance of each Gaussian was not permitted to drop below a certain minimum threshold. It was found that without this rule, the Gaussians' variance would become very small and as a result it did not account well for camera noise.

A drawback of this background modelling approach is that it models each pixel separately and does not take the scene context into account. For example, if the colour of a person's clothing is very similar to the background that it is occluding it will not be detected as foreground. Methods to overcome this limitation might involve using information about previously tracked objects and other high-level reasoning techniques. However, the low-level approach we take is to notice that although the foreground and background may have identical colours in the visible spectrum, they may have features in other spectral bands that can discriminate between them. Combining infrared segmentation with the foreground detection, as will be explained shortly, produces much cleaner and more accurate results. After each pixel is classified as background, shadow or foreground, some morphological cleaning is performed to remove noise.

3.2. Infrared Background Modelling

The infrared background is modelled using a simple averaging of frames method, using the expected sufficient statistics equations and then switching to the L-recent window update equations when the first L frames have been processed, as described in [7] and equation 1.

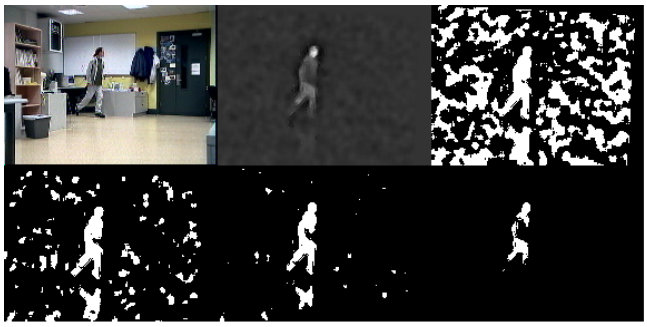

Fig. 3. Visible Image, Filtered Infrared Image, Filtered image thresholded at different thresholds: 0, 1, 2, 8

$$B^t_{x,y} = \begin{cases} I^0_{x,y} & \text{if } t \leq 1 \\ (\frac{t-1}{t})(B^{t-1}_{x,y}) + (\frac{1}{t})(I^{t-1}_{x,y}) & \text{if } 1 < t < L \\ (\frac{L-1}{L})(B^{t-1}_{x,y}) + (\frac{1}{L})(I^{t-1}_{x,y}) & \text{if } t \geq L \end{cases} \quad (1)$$

where $B^t_{x,y}$ is the background model of the pixel at location (x, y) at time t and $I^t_{x,y}$ is the pixel at location (x, y) of the frame at time t. L is a parameter that controls how quickly a new object is incorporated into the background model. The infrared foreground is calculated by simply subtracting the background model from the current frame, filtering the result and thresholding, as described by the equation:

$$F^t_{x,y} = M_T(A(I^t_{x,y} - B^t_{x,y})) \quad (2)$$

where F^t is a binary image representing the detected foreground of the infrared band at time t, M is a thresholding function returning a binary image using the threshold T and A is an adaptive filter described in the next section.

3.3. Detection of Hot Regions

Due to the nature of infrared imaging and to the absorption of the glass beam-splitter, the infrared foreground images obtained from the rig are very noisy and have low contrast. The first step is to remove as much noise as possible while preserving the important details. Gaussian smoothing [9] is effective at noise removal but it also blurs edges and removes the finer details. In our approach, we use adaptive filtering which takes the edge orientation into account and performs smoothing along the edge but not perpendicular to

it. Firstly, the magnitude and direction of edges are computed by smoothing the image with a Gaussian mask and then calculating the gradients in the x and y direction. If the magnitude of the edge at a pixel is below a certain threshold, it is smoothed with a traditional Gaussian filter. Otherwise, the pixel is smoothed with a directional Gaussian mask, oriented along the pixel's edge direction. The filtered image must then be thresholded to detect regions that are warmer than the background noise. A threshold value, T, classifies pixels with a value below T as background, otherwise they are foreground. As can be seen in Figure 3, a low threshold will include too much noise but a higher threshold will lose the fine segmentation details and will remove the colder and more insulated parts of the person, usually the legs and lower torso.

There are various approaches to choosing the correct threshold. A constant threshold may work well for some test images but it may not work well if the scene changes. A dynamic threshold could be calculated based on image features. However, an effective threshold selection process should take advantage of the information available from the visible image. The foreground pixels extracted from the visible image provide information that allows us to judge the similarity between our infrared thresholding and the visible foreground detection. Using the idea of mutual information, a ratio can be defined to measure the agreement between the infrared and visible spectrum foreground detection. Thus, we wish to choose a value of T that will maximise the ratio. We define two ratios:

$$R_1 = \frac{P_{(1,1)}}{P_{(1,0)}+P_{(0,1)}} \tag{3}$$

$$R_2 = \frac{P_{(0,0)}+P_{(1,1)}}{P_{(1,0)}+P_{(0,1)}} \tag{4}$$

where $P_{(x,y)}$ is the total sum of pixels whose visible classification is x and whose infrared classification is y. For example, $P_{(0,1)}$ is the total number of pixels who are classified as background by the visible analysis and as foreground by the infrared analysis. Therefore, R_1 is the ratio between agreed foreground pixels and total disagreed pixels. R_2 is the ratio of total agreed pixels to total disagreed pixels. The foreground selection threshold, T, for the infrared image is chosen so as to maximise R_1. If the value of the ratio is less than 0.1 then T is chosen so as to maximise R_2. This is because R_1 puts more emphasis on agreeing on foreground pixels and can cause the threshold to drop very low to *agree* with the foreground noise when there are no objects in the scene.

After the threshold selection, a binary image is created so that each pixel has value one if it is classified as foreground by the visible or infrared analysis, zero otherwise. All connected regions that do not contain at least one visible foreground pixel and an infrared foreground pixel are removed. Regions with a small number of pixels are considered noise and are also removed.

4. RESULTS

The video sequences used in our experiments have a resolution of 192x144 and were captured at 25 frames per second. Figure 4 shows a typical example of where techniques such as background modelling and motion analysis would have severe difficulties when considering only visible spectrum video. Multimodal analysis is able to provide robust segmentation of the person in this case.

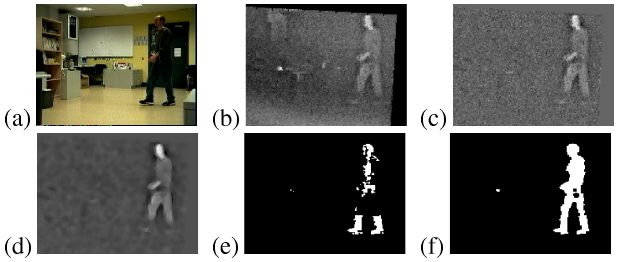

Fig. 4. (a) visible image. (b) aligned infrared image. (c) background subtracted infrared. (d) adaptive filtering on infrared. (e) foreground detected in visible domain. (f) visible and infrared fusion.

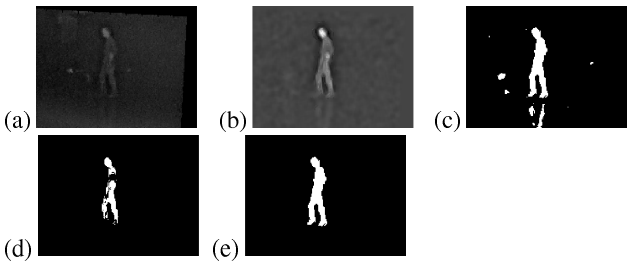

Fig. 5. Segmentation using infrared only: (a) aligned image. (b) filtered background subtracted image. (c) thresholding with T_L. (d) thresholding with T_H. (e) combined hysteresis segmentation.

4.1. Changing Lighting Conditions

One particular scenario that interested us was to investigate how a system could cope without any relevant data from the visible video stream. Visible background modelling often fails due to abrupt changes in lighting conditions, such as lights being dimmed or turned on/off, as well as changes in ambient lighting. We detect these failures of the background model in the visible spectrum by detecting when 80% of the image is classified as foreground, similarly to [10]. Previously, our detection of hot regions was based on using the visible information for threshold selection and seeding the resulting regions with the foreground pixels. Without the visible signal, we set a low threshold, T_L, and we opt to use only regions that contain 'very hot' pixels. These usually correspond to exposed skin regions such as the head and forearms. Pixels are classified as 'very hot' if their value (in the infrared background subtracted image) is greater than a high threshold, T_H. This is essentially a hysteresis [9] segmentation using the high and low thresholds. An example of this can be seen in Figure 5.

4.2. Intruder Detection

To simulate a real surveillance application, we defined a restricted area and aimed to detect whether a person had entered this area. To establish the likelihood of a given object being inside the area, a bounding box was defined and an alarm measure was calculated.

$$\text{probability of intruder} = \frac{overlap(O_k, B)}{area(O_k)} \tag{5}$$

where O_k is the bounding box of object k in the frame and B is the bounding box of the restricted area. The resulting signal is

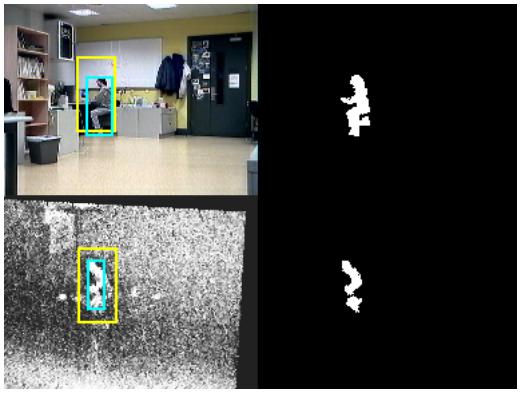

Fig. 6. Intruder detection; Top row: restricted area in yellow, object bounding box in blue, detected object on right. Bottom row: detection in total darkness. Histogram equalised infrared image on left, detected object on right

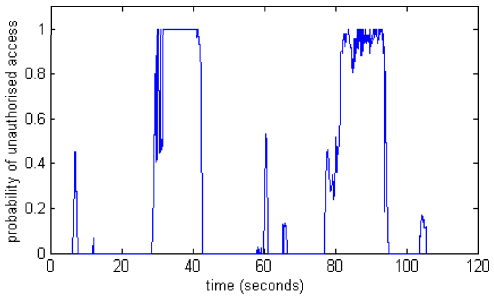

Fig. 7. Typical example of intruder detection

median filtered to reduce the effects of spurious noise. Figure 6 shows two examples of detection. Since the output is a probability and not a binary value, it is possible to control the sensitivity of the alarm. In our application, it was found that checking if the alarm measure was above 0.8 for 5 seconds was a reliable method of intruder detection. Two examples of unauthorized access are detected in the data shown in Figure 7 between 30 and 42 seconds and between 85 and 93 seconds. The first detection is achieved in total darkness. The (non-intruder) spikes in the graphs were caused by people walking in front of the restricted area, occluding it, and therefore their bounding box would partially overlap with the restricted area.

5. CONCLUSIONS AND FUTURE WORK

This paper has demonstrated the advantages of multi-modal video analysis for object segmentation, both in terms of more accurate video object extraction and in its improved robustness to lighting changes. Short-term future work will focus on the refinement of object boundaries using edges in the visible spectrum. Face and skin detection is another low-level module that could be made significantly more robust by combining infrared and visible analysis since exposed skin and especially the human head area, emit significant amount of infrared radiation.

Long-term work will focus on incorporating additional input devices into the hardware rig. The challenge of building a multi-modal analysis system involves determining the optimum method of combining the analyses of the individual modalities, so as to utilise the strengths of each one while remaining robust to failures of some of the modalities. Besides visible and infrared information, knowledge of depth is the next modality we will investigate. We plan to add a second CCTV camera to the rig and to use stereo-vision techniques to incorporate depth information, which is useful both in the boundary detection of objects and also in object tracking during occlusion.

6. ACKNOWLEDGEMENTS

This material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/I361 and sponsored by a scholarship from the Irish Research Council for Science, Engineering and Technology (IRCSET): Funded by the National Development Plan. The authors would also like to express their gratitude to Mitsubishi Electric Research Labs (MERL) for their contribution to this work.

7. REFERENCES

[1] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics*, 34(3):334–350, August 2004.

[2] F. Xu and K. Fujimura. Pedestrian detection and tracking with night vision. In *Procs. IEEE Intelligent Vehicles Symposium*, June 2002.

[3] M. Bertozzi, A. Broggi, T. Graf, P. Grisleri, and M. Meinecke. Pedestrian detection in infrared images. In *Procs. IEEE Intelligent Vehicles Symposium*, pages 662–667, June 2003.

[4] Shih-Schn Lin. Review: Extending visible band computer vision techniques to infrared band images. Technical report, GRASP Laboratory, Computer and Information Science Department, University of Pennsylvania, 2001.

[5] M. Irani and P. Anandan. Robust multi-sensor image alignment. In *International Conference on Computer Vision*, pages 959–966, 1998.

[6] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2003.

[7] P. KaewTraKulPong and R. Bowden. An improved adaptive background mixture model for real-time tracking with shadow detection. In *2nd European Workshop on Advanced Video-based Surveillance Systems, Kingston upon Thames*, 2001.

[8] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of CVPR99*, pages II:246–252, 1999.

[9] M. Sonka, R. Boyle, and V. Hlavac. *Image Processing, Analysis and Machine Vision*. PWS, November 1998.

[10] I.Haritaoglu, D.Harwood, and L.Davis. Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (22):781–796, August 2000.