

# MULTISPECTRAL OBJECT SEGMENTATION AND RETRIEVAL IN SURVEILLANCE VIDEO

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

Centre for Digital Video Processing, Dublin City University, Ireland

## ABSTRACT

This paper describes a system for object segmentation and feature extraction for surveillance video. Segmentation is performed by a dynamic vision system that fuses information from thermal infrared video with standard CCTV video in order to detect and track objects. Separate background modelling in each modality and dynamic mutual information based thresholding are used to provide initial foreground candidates for tracking. The belief in the validity of these candidates is ascertained using knowledge of foreground pixels and temporal linking of candidates. The Transferable Belief Model [1] is used to combine these sources of information and segment objects. Extracted objects are subsequently tracked using adaptive thermo-visual appearance models. In order to facilitate search and classification of objects in large archives, retrieval features from both modalities are extracted for tracked objects. Overall system performance is demonstrated in a simple retrieval scenario.

**Keywords:** Infrared surveillance, Infrared image sensors, Information retrieval, Video signal processing, Tracking.

## 1. INTRODUCTION

World events have ensured that security and surveillance have received much attention in recent years. The desire to partially or fully automate many of the mundane tasks involved in observing or reviewing hours of surveillance video has stimulated research in computer vision techniques for visual surveillance. In this area, the fusion of multiple sources of information (e.g. from multiple cameras of different modalities) is a challenging task, but one that has the potential to provide for more robust systems by leveraging the combined benefits of using different modalities whilst compensating for failures in individual modalities.

In our work, we investigate the advantages of capturing thermal infrared video in parallel with standard visible spec-

trum CCTV. In building our surveillance system, the individual advantages of each modality are retained, such as colour-based analysis (using the visible spectrum) and 24-hour operation without external lighting (using thermal infrared). Other ancillary advantages arise when the combination of modalities is considered. An example is the possibility of extracting a *signature* of an object in each modality, which indicates how useful each modality is in tracking that object, as well as providing features that can be used for classification or retrieval. Another example is improved robustness against camouflage, as foreground objects are less likely to be of a similar colour and temperature to the background.

This paper is organised as follows: In section 2, we provide a brief background literature review to contextualise our work. Section 3 describes our detection and tracking vision system and the features extracted from the tracked objects. We present results in section 4, showing illustrative tracking results and retrieval experiments and give our conclusions and directions for future work in section 5.

## 2. BACKGROUND

Given the large amount of research in the area of visual surveillance in recent years, in this section we only consider those works that help contextualise and motivate our approach. In [2], Cucchiara argues that the integration of video technology with sensors and other media streams will constitute the fundamental infrastructure for new generations of multimedia surveillance systems, thereby providing us with the motivation for investigating the benefits of combined CCTV/infrared analysis. In [3], Hu *et al* review the various approaches to designing surveillance-system components and identify a common processing framework used in the vast majority of systems. Our proposed system conforms to this framework. In [4], a real-time surveillance system is described that performs analysis on gray-scale or thermal video, to detect and track people and to identify human body parts and activities in an outdoor environment. However, unlike the approach in [4] we use multiple spectral bands and fuse the information to detect and track objects.

---

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.

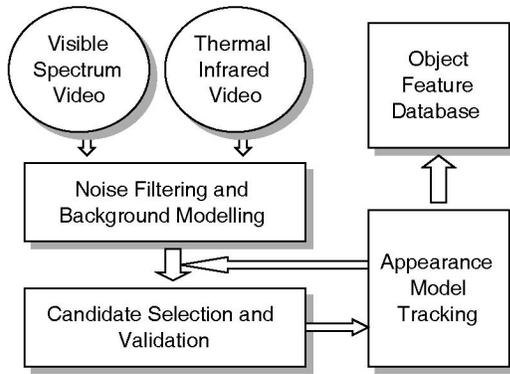


Fig. 1. System Block Diagram

### 3. SYSTEM ARCHITECTURE

Figure 1 shows the overall surveillance system architecture. In the following sections, the individual processing blocks are explained in more detail. The camera rig set-up that ensures rectification between visible and infrared camera feeds was previously described in [5].

#### 3.1. Noise Filtering and Background Modelling

To reduce the noise in the thermal images, structure-preserving SUSAN noise filtering is used [6]. The background model for each modality (thermal infrared and visible spectrum) are modelled using the approach in [7] – an improved version of the approach proposed by Stauffer and Grimson [8] that incorporates faster initialisation and shadow removal. Note that the thermal images only have one band, whereas the CCTV images are RGB colour. Since there is no correlation in brightness values between thermal and visible [9], we can model the modalities separately so that failure in one can be compensated by the other. The background models return a difference map indicating, for each pixel, the number of standard deviations from the mean of the nearest matched Gaussian background model. This is then thresholded to produce a foreground map. We determine the decision thresholds using mutual information thresholding [10]. If the mutual information quality score is low, we revert to using the standard threshold of 2.5 standard deviations.

#### 3.2. Candidate Selection and Validation

We use three foreground maps (visible, infrared, visible OR infrared) and first remove all pixels that can be explained by objects currently being tracked. Tracking candidates are obtained by fitting bounding boxes to connected component regions in these foreground maps. In each frame, every previous candidate is linked to one current candidate. A similarity measure between past and current candidates is computed as the area of the larger candidate divided by the overlap area.

The most similar candidates are linked first, until no further links can be made. Validity beliefs (see below), indicating whether or not a candidate should be tracked, are then updated for all linked candidates, using the number of foreground pixels inside the current candidate and the candidate-link similarity measure. New candidates are allocated significant doubt in their validity beliefs, as there is not enough evidence to support their removal or their tracking validity. After each frame is processed, candidates that have high invalidity belief are removed.

The Transferable Belief Model (TBM) [1] is a framework for expressing degrees of belief and allowing evidence from various sources to be combined. It also provides a powerful framework for uncertainty (or doubt) to be expressed. Given some evidence, it allows the construction of a basic belief assignment function  $m : \Omega \rightarrow [0, 1]$ , allocating a belief mass to each possibility, such that  $\sum m(A) = 1$ , when the summation is over all possibilities. In our scenario, we examine tracking candidates individually and have only three possible beliefs: YES (the candidate should be tracked, denoted  $Y$ ), NO (it should not be tracked, denoted  $N$ ), YES OR NO (the decision is in doubt, denoted  $D$ ). Our evidence for tracking validity comes from the number of foreground pixels within the candidate area and the similarity measure between it and the previous candidate. These values are mapped to fuzzy belief assignment functions and are fused to compute the current candidate's beliefs. They are then updated by fusing them with the beliefs of the previous candidate. Belief fusion is done as follows:

$$m(Y) = m_1(Y)m_2(Y) + m_1(Y)m_2(D) + m_2(Y)m_1(D) \quad (1)$$

$$m(N) = m_1(N)m_2(N) + m_1(N)m_2(D) + m_2(N)m_1(D) \quad (2)$$

$$m(D) = m_1(D)m_2(D) + m_1(Y)m_2(N) + m_2(Y)m_1(N) \quad (3)$$

where  $D = Y \cup N$ ,  $m_1$  and  $m_2$  are the belief assignments of the sources to be fused. The two pieces of evidence being fused may give conflicting information and this is added to the doubt via the last two terms in equation 3. Thus, if there is disagreement between sources, there will be more uncertainty and the decision to track will be postponed until sufficient belief exists to remove the candidate or to initiate a tracker. After initiating a tracker, all overlapping candidates are removed.

#### 3.3. Appearance Model Tracking

Given a decision to track, we model the object to be tracked using the adaptive appearance model used in our previous work [11], where we evaluated various fusion strategies for tracking. Inspired by the work in [12], we use object models represented as rectangular grids of  $d$  pixels, with each pixel modelled as a Gaussian distribution. Additionally, each pixel is assigned an importance depending on how often it occurs as foreground. This is used to down-weight the background pixels that should not be tracked. Unlike [12] where image

patches are matched using maximum likelihood probability, we use a separate model for each modality and fuse their matching scores. We found that this tracking approach outperformed other fusion strategies [11]. The object similarity measure we use is given by:

$$S(X) = \frac{\sum_{j=1}^d I(j)P(X_j|\theta(j))}{\sum_{j=1}^d I(j)} \quad (4)$$

where  $I$  is the pixel importance,  $X$  is the image patch being examined and  $\theta$  represents the Gaussian model for the pixel. The model parameters are updated as in [12] using an update parameter  $\alpha = 0.05$ . The appearance models are used to *explain* (and remove) foreground pixels if the model determines the importance value of the pixel to exceed 0.5.

### 3.4. Feature Extraction

Given a segmented and tracked object, it is possible to extract useful features for each frame. The features currently extracted are: (a) coordinates of a rectangle tightly enclosing the object, (b) the size in pixels of the object (from the appearance model), (c) the number of foreground object pixels detected in both the infrared and visible spectrum. The latter features indicate the object's thermal and visible *signatures* and are useful in object classification, as well as for providing the system with feedback on which modality is generating the most reliable tracking information (note this last item is not currently not exploited in our system, but is left for future work). Although the extracted features are simple, more useful features can be derived from them. We found that the height-to-width ratio is useful for general classification. Also, by modelling the ground plane, assuming that each object is approximately vertical and its lowest point touches the ground, it is possible to obtain relative physical height and width measurements.

## 4. EXPERIMENTAL RESULTS

### 4.1. Segmentation and Tracking

In figure 2, we show illustrative results from our object extraction system. The corresponding thermal infrared images are shown in the right-hand column. Our multimodal video sequences were captured at various times of the day and night, over-looking a busy area of our university. The objects that were extracted include people, cyclists, motorcyclists and cars.

Figure 3 shows some examples of when tracking can fail. In the first example, three people enter the scene in close proximity and are subsequently tracked as two objects. Similarly, in the second example, two people are tracked using a single tracker, as they are repeatedly detected as one object in the foreground maps.

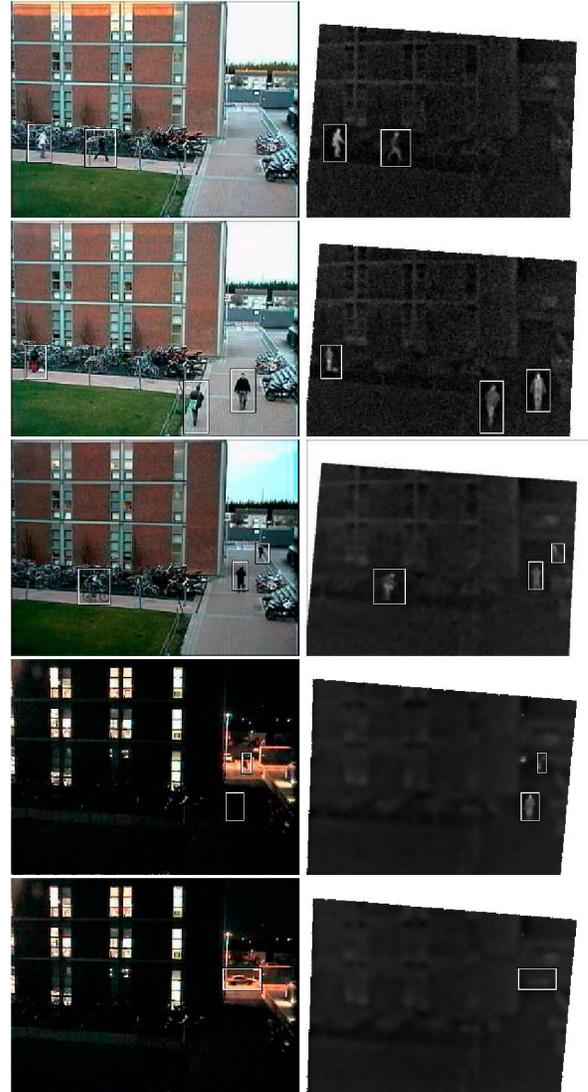


Fig. 2. Sample results for daytime and nighttime video of the same scene

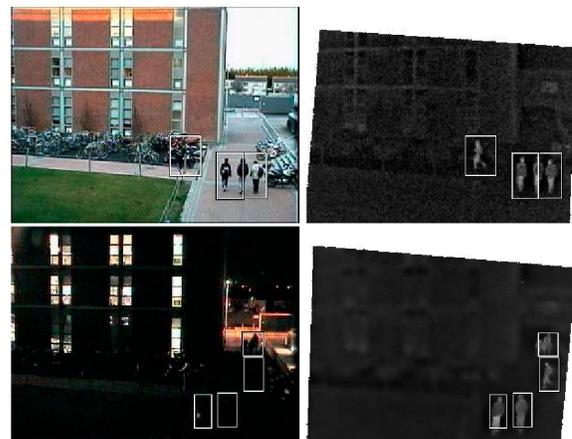


Fig. 3. Examples of tracking failures

## 4.2. Object-based Retrieval

In order to demonstrate the potential usefulness of our system in a real scenario featuring large volumes of surveillance video that it is required to search, we have developed a rudimentary retrieval system based on the extracted objects and features. Our database includes features from 390 objects, extracted every frame for an average of 212 frames per object. For each object, we reduced each feature to a single value by performing an averaging, weighted by the object size. We used only the following features: height to width ratio, thermal and visible signatures, and physical size (computed using the ground plane assumption). These were found to be the most discriminative features. We then calculated the Mahalanobis distance between each pair of objects to perform object matching. Figures 4 and 5 show the top 15 most similar objects to the target object (in top left corner of each figure), for the examples of searching for cars and people, respectively. In each case, the corresponding infrared images are shown below. Black areas in the infrared images indicate the infrared image boundary, as the two modalities did not overlap completely after alignment.

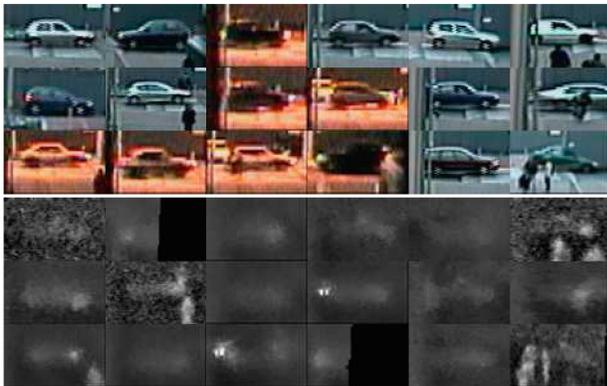


Fig. 4. Retrieved cars using object-object similarity



Fig. 5. Retrieved people using object-object similarity

## 5. CONCLUSION AND FUTURE WORK

In this paper we presented an object segmentation and tracking system that fuses CCTV and infrared analysis and facilitates feature extraction for search and retrieval from archives of surveillance video. We provided illustrative results for both. Currently, we do not use colour to track objects but our model can be easily extended to handle extra tracking information, such as RGB colour. In this case, the assumption of feature independence does not hold, but this can be overcome by switching to a less correlated colourspace (such as the CIE Lab colourspace) or by increasing the computational complexity and using a full covariance matrix. This is targeted as future work. Future work will also involve increasing the size of our object database by capturing and processing additional multimodal video sequences using our system in multiple different locations. We will also investigate on-line learning using the extracted features to train classifiers to identify objects.

## 6. REFERENCES

- [1] P. Smets, "The combination of evidence in the transferable belief model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 447–458, 1990.
- [2] R. Cucchiara, "Multimedia surveillance systems," in *VSSN 05: Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, New York, NY, USA, 2005, pp. 3–10.
- [3] 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*, vol. 34, no. 3, pp. 334–350, August 2004.
- [4] I. Haritaoglu, D. Harwood, and L. Davis, "Real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 22, pp. 781–796, August 2000.
- [5] C. Ó Conaire, E. Cooke, N. O'Connor, N. Murphy, and A. F. Smeaton, "Fusion of infrared and visible spectrum video for indoor surveillance," in *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Montreux, Switzerland, April 2005.
- [6] S. M. Smith and J. M. Brady, "Susan - a new approach to low level image processing," *Int. Journal of Computer Vision*, vol. 23, no. 1, pp. 45–78, May 1997.
- [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*, 1999, pp. II:246–252.
- [9] M. Irani and P. Anandan, "Robust multi-sensor image alignment," in *International Conference on Computer Vision*, 1998, pp. 959–966.
- [10] C. Ó Conaire, N. O'Connor, E. Cooke, and A. Smeaton, "Detection thresholding using mutual information," in *VISAPP: International Conference on Computer Vision Theory and Applications*, Setúbal, Portugal, Feb 2006.
- [11] C. Ó Conaire, N. E. O'Connor, E. Cooke, and A. F. Smeaton, "Comparison of fusion methods for thermo-visual surveillance tracking," in *International Conference on Information Fusion*, July 2006.
- [12] S. Zhou, R. Chellappa, and B. Moghaddam, "Appearance tracking using adaptive models in a particle filter," in *Proc. of 6th Asian Conference on Computer Vision (ACCV)*, Jan 2004.