Background Modelling in Infrared and Visible Spectrum Video for People Tracking

C. Ó Conaire, E. Cooke, N. O'Connor, N. Murphy, A. Smeaton
Centre for Digital Video Processing, Adaptive Information Cluster, Dublin City University, Ireland

Abstract

In this paper, we present our approach to robust background modelling which combines visible and thermal infrared spectrum data. Our work is based on the non-parametric background model described in [1]. We use a pedestrian detection module to prevent erroneous data from becoming part of the background model and this allows us to initialise our background model, even in the presence of foreground objects. Visible and infrared features are used to remove incorrectly detected foreground regions, allowing our model to quickly recover from ghost regions and rapid lighting changes. An object-based shadow detector also improves our algorithm's performance.

1. Introduction

Background estimation is a primary module in many vision tasks and is used to distinguish important objects from the normal background scene. In this paper, we present our approach to robust background modelling which combines visible and thermal infrared spectrum data.

The benefits we obtain from using infrared data, as well as visible information, are twofold. Firstly, as a complementary modality to the visible spectrum, detection can be improved by using the strengths of each medium. Infrared detection aids visible analysis in low-lighting conditions and when an object has similar colour to the background. Therefore, it has the capability to operate on a 24-hour basis, as thermal infrared video detects primarily emitted radiation, thus does not require daylight to function. Visible spectrum detection has a higher resolution and can detect objects whose temperature is not significantly different from the background. Secondly, the infrared data provides extra features that can be used alongside size and edge features to detect errors in the foreground regions and quickly correct the background model. We use a pixel-based likelihood map to accumulate evidence that a foreground pixel should become part of the background. In the initialisation phase, we use a pedestrian detection module to prevent erroneous data from becoming part of the background model, allowing our model to initialise correctly, even in the presence of foreground objects.

Our paper is structured as follows: the following section describes previous work in the areas of background modelling and the use of thermal infrared video. The third section describes our algorithm for robust foreground detection. We show some examplary results in section four and discuss these results and future work in section five.

2. Literature Review

2.1. Background Modelling

Background modelling assumes that the video scene is composed of a relatively static model of the background, which becomes partially occluded by objects that enter the scene. These objects (usually people or vehicles) are assumed to have features that differ significantly from those of the background model (their colour, thermal or edge features, for example). Calculating a distance measure between the current scene and the modelled background is known as foreground extraction and is usually used to produce a binary foreground image.

Reliable background modelling is difficult to achieve in certain scenarios. For example, in a crowded room with many people, the background may only ever be partially visible. Another problematic scenario is in a scene with low levels of lighting, such as a night-time scene with only street lighting, or a scene with varying degrees of lighting, such as when part of the scene is in shadow and the other part is sunlit. The movement (or apparent movement) of background objects is problematic too. Examples of this include moving trees and vegetation, flickering computer or TV screens, flags or banners blowing in the wind, etc.

Background modelling is a very active research area and has progressed considerably from the early 'subtract and threshold' approaches. The algorithm described in [2] models each pixel as a sum of K Gaussian distributions in RGB space ($1 \leq K \leq 5$). Each pixel's background model is updated continuously, using online estimation of the parameters. In [3], an illumination independant background model is described that uses a correlation measure between image blocks to detect foreground regions. An object-based foreground extraction scheme is described in [4]. The algorithm used in the W^4 [5] system works on monochrome video and initiates training periods where, for each pixel,

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

its maximum value, minimum value and largest interframe difference are calculated. These values are used as the background model and the foreground is determined by straight-forward thresholding and morphological processing. A good review of the most commonly cited background modelling techniques can be found in [6].

We adopt the non-parametic background model described in [1]. The authors of [7] cite this model as superior to [3] and [2], although its storage requirements may be too great for some systems. We decided to use this model, as it is easily extended to handle the additional thermal infrared band. Their method is decribed in the section on background modelling.

2.2. Infrared Imaging

Thermal infrared imaging has long remained a relatively small academic research area due to the cost of the thermal imaging systems, the low resolution and the noisy images that are produced. Thankfully, the decreasing costs and improved optics are permitting more researchers to enter the field. Thermal imaging is a complementary technology to visual imaging, as it relies on emitted, rather than reflected radiation. Also, it is of great benefit to monitoring and surveillance systems, as it can operate on a 24 hour basis and is most reliable at detecting and tracking hot objects, such as people and vehicles, which are normally the primary objects of interest in surveillance. In [8] the application of traditional image analysis techniques to infrared is discussed, as well as conducting a good review of infrared imaging research. Pedestrian detection approaches using thermal imaging are described in [9] and [10].

To obtain our video test data, we use the camera system described in [11] which allows simultaneous capture of infrared and visible spectrum video. Temporal alignment is achieved using the cameras' *gen-lock* input which allows their frame clocks to be synchronised. Spatial alignment is achieved using a planar homography, whose parameters are calculated by manually selecting numerous corresponding points in both modalities and computing the homography with least mean squared error.

3. Algorithm

Figure 1 shows block-diagrams of our main system components.

3.1. Diffusion Filtering

Infrared images contain high noise due to the nature of infrared radiation, since every hot body, including the imaging device itself, emits non-insignificant amounts of radiation at this wavelength. We reduce this noise using anisotropic diffusion [12] that inhibits smoothing at edges, thus greatly reducing the blurring of edges that results from Gaussian

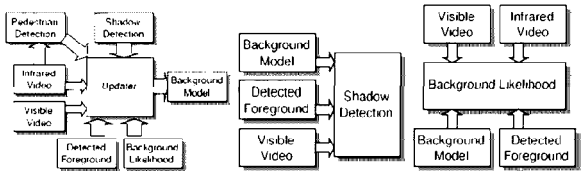

Figure 1: System Schematic: (a) Background updating, (b) Shadow detection and (c) Likelihood map calculation

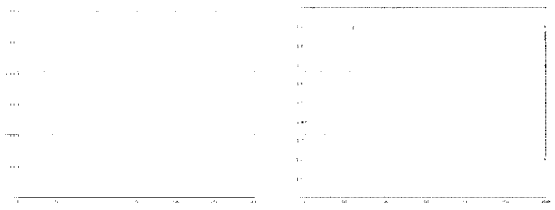

Figure 2: (a) Typical infrared histogram (b) Associated importance function

smoothing. Iterative filtering is performed on an infrared image, I, by repeating the following steps until convergence:

1. Calculate the magnitude of the gradient of a smoothened version of I, $M = |\nabla G(I)|$, where G is an isotropic Gaussian smoothing kernel.
2. Calculate coefficients for filtering, $C = \frac{1}{M+1}$
3. Multiply each pixel in I by its corresponding pixel in C
4. Set $I = F(I)/F(C)$, where $F(X)$ is equivalent to X when each pixel is replaced by itself plus the sum of its eight neighbours.

3.2. Pedestrian Detection

We perform a rough pedestrian detection in the infrared image by first segmenting the image into regions and then discard non-pedestrian regions using size, aspect ratio and thermal features. From observation of typical infrared imagery, we note that the histogram invariably contains a dominant Gaussian distribution which represents the ambient temperature of the environment and is essentially noise, such as shown in figure 2(a). This makes sense, as thermal noise has a Gaussian distribution [12]. Relevant objects, such as people, usually consist of brighter pixels that lie far outside this distribution. Using the histogram, we create an importance score (Figure 2(b)) for each brightness

value and then replace each pixel with the importance value of its brightness. We make the assumption that pedestrians are warmer than the environment, so we replace with zero all pixels whose brightness is less than the mean brightness. The importance function is calculated as follows:

$$u(x) = \left(\frac{1}{\frac{h(x)}{max(h)} + 1} \right)^n \quad (1)$$

where x is a pixel brightness, $h(x)$ is the histogram of the image and n is a parameter which deemphasises values inside the noise distribution. Experimentation has shown that n=10 is a suitable value however slight deviations either side of this value do not significantly alter results. After replacing each pixel by its importance value, we perform a hysteresis segmentation, with a lower and upper threshold, T_L and T_U, to obtain regions. In other words, all pixels with values less than T_L are discarded, then of all remaining connected-components, only those containing at least one pixel with a value greater than T_U are considered valid regions.

For each region, we extract the following features: s_r, its area in pixels, b_r, the height-to-width ratio of its bounding box and, using brightness values from the original infrared image (not the importance image), a_r, its average brightness and m_r, its maximum brightness. Regions are classified as non-pedestrians if any of these statements are true:

1. $s_r < S_{min}$ or $s_r > S_{max}$
2. $b_r < 1$ or $b_r > 5$
3. $(a_r - m)/\sigma \leq 2.5$
4. $(m_r - m)/\sigma \leq 4$

where m and σ are the mean and standard deviation of the pixel brightness in the infrared image and S_{min} and S_{max} are thresholds on the minimum and maximum pedestrian sizes. We noted that occasionally, pedestrians would overlap each other or be joined to noise. Thus non-pedestrian regions are classified as pedestrians if $(m_r - m)/\sigma > 6$, as human skin is usually significantly hotter than other objects. Finally, the resulting binary image is dilated with a large structuring element to ensure no parts of pedestrians are missed.

The goal of this detection module is not to miss any pedestrians, so that their pixels are not put into the background model. Therefore the thresholds are set to obtain a low false negative rate. The output of the pedestrian detection is a binary mask, $P(x, y)$, for the image. For our experiments, S_{min} was set to approximately one quarter the size of the smallest expected pedestrian and S_{max} was set slightly greater than the larger expected pedestrian size.

3.3. Background Model

Our background model is based on the non-parametric model described in [1]. For each pixel, the model stores N samples that are assumed to belong to the background distribution. For a new pixel x_t, the probability that it came from the background distribution is:

$$Pr(x_t) = \frac{1}{N} \sum_{i=1}^{N} \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{1}{2}\frac{(x_{t_j} - x_{i_j})^2}{\sigma_j^2}} \quad (2)$$

where d is the number of bands in the image. In our case, we have 4 bands: L, U, V and infrared. These bands are less correlated than RGB and thus the assumptions of independence are more valid in the model. The variance for each band, σ_j^2, is obtained by calculating the median, q, of $|x_i - x_{i+1}|$ for each consecutive pair in the sample and setting $\sigma = \frac{q}{0.68\sqrt{2}}$. A new pixel is classified as foreground if its probability of belonging to its background model is less than a threshold value. Some morphological operations are then used to remove pixel noise and to close holes in foreground regions.

We extend this model by allowing a pixel to have an *unknown* value; essentially to have zero samples in its background model. This is useful in the case where we detect that a pedestrian is occluding a background pixel at this point and do not want to update the model with erroneous data. A new pixel is classified as foreground if its background model is unknown. Background updating is done for two reasons. Firstly, to cope with gradual changes in lighting, such as the change from day to night, and secondly, to place objects in the background that are determined to have a high likelihood to be part of it; either objects that have remained static for some time or incorrectly labelled foreground regions, such as those caused by rapid changes in brightness. The updating procedure must carefully avoid placing parts of foreground objects in the model, while also being able to quickly adapt and perform rapid updates in the presence of ghosts or fast lighting changes. Our background model is updated continuously, thus it can adapt to gradual changes in lighting.

In our initialisation phase, each pixel is added to the background model if it is not detected as part of a pedestrian. In normal operation, the background, B, is updated in the following manner:

1. if $\Gamma(x, y) \geq 1$, update $B(x, y)$, where Γ is described in subsection 3.5.
2. otherwise, if $B(x, y) =?$ (unknown), update $B(x, y)$ only if $P(x, y) = 0$
3. otherwise, if this pixel is classified as shadow, it is not updated

4. otherwise, update this pixel if it has been detected as background

For a given pixel, if its background model already has N samples, then updating its background model consists of discarding its oldest sample and inserting the current pixel. The variance values are also updated periodically.

3.4. Shadow Detection

Shadows are first crudely detected in each frame by calculating the decrease in brightness and the chromaticity change for each pixel. This is somewhat similar to shadow detection in [1] but we use only the mean values of each band and not multiple samples. A binary image is created by classifying pixels within certain bounds as potential shadow pixels. All connected-component regions with an area less than T, a size threshold, were discarded. A pixel, p, is a potential shadow pixel if the following conditions all hold true:

1. $0.8 \leq C_L \leq 0.98$, where the change in Luminance, $C_L = \frac{p_L}{B_L}$
2. $C_{UV} \leq 20$, where the change in Chrominance, $C_{UV} = \sqrt{(B_U - p_U)^2 + (B_V - p_V)^2}$

where B_L, B_U and B_V are the average background values for each band. We then compute the detected foreground regions that overlap with the potential shadow regions and calculate $\theta = |S|/|F|$, where $|S|$ is the area of the shadow and $|F|$ is the area of the foreground region. True shadows are determined as those where $\alpha \leq \theta \leq \beta$, α and β being parameters determining the allowed size of a shadow for a given object size. This object based approach is quite robust and similarly to [4], associates shadows with detected objects.

3.5. Background Likelihood

Foreground objects, such as cars, should be expected to become background if they remain in the same position for a long period of time. Also, in terms of invalid foreground regions, if there is enough evidence to suggest that they are actually caused by background model errors or ghosts, the background should be updated to include them. We introduce a likelihood image, Γ, that accumulates evidence for each pixel that it should belong to the background model. Four features contribute to this likelihood:

1. Time: if an object remains in the same position long enough, it should become part of the background.
2. Size: small foreground regions are more likely to be caused by background errors.
3. Edge Magnitude: as discussed in [13], edges provide evidence that a foreground region is a ghost.
4. Thermal brightness: colder objects are more likely to be part of the background.

Therefore, at each frame, for every pixel, (x, y), detected as foreground, Γ is updated as follows:

$$\Gamma_{x,y,t} = \Gamma_{x,y,t-1} + C_T + f_{T_s,\sigma_s}(s_r) + f_{T_e,\sigma_e}(e_r) + f_{T_b,\sigma_b}(b) \quad (3)$$

$$f_{T,\sigma}(x) = C(1 - \frac{1}{1 + e^{\frac{-(x-T)}{\sigma}}}) \quad (4)$$

$$e_r = \frac{1}{|\delta r|} \sum_{i \in \delta r} (|\nabla I(i)| - |\nabla B(i)|) \quad (5)$$

where s_r is the size of the foreground region containing (x, y), b is the maximum thermal brightness in a 7x7 window around (x, y) that is also within the same foreground region. b is normalised by subtracting the mean brightness and dividing by the standard deviation. C_T and C are constants controlling how fast a static object becomes background. $f_{T,\sigma}$ is based on a sigmoid function centred at T whose transition width is controlled by σ. δr is the set of border pixels of region r. By using the sigmoid function, our thresholds do not have to be hard and instead provide a smooth output for small changes in feature values. Parameters $T_s, \sigma_s, T_e, \sigma_e, T_b, \sigma_b$ were chosen empirically. In equation 5, we provide a similar measure of border edge strength as used in [13]. Their approach was to detect which edges in the borders of foreground region were different from the background model edges and then to find the average magnitude of these edges in the current image. Our metric does not require the detection of changed edges but uses the difference between the current edges and the background edges on the border to provide evidence of background likelihood, thus removing the need for am empirically-chosen detection threshold.

4. Results

Our pedestrian detection module performs very well on our dataset and on most images from the OTCBVS test-set, as shown in figure 3. The algorithm failed on one sequence in the OTCBVS test-set, figure 3(c), where our assumption that pedestrians would appear brighter than the background did not hold.

Figure 4 shows an example of our object based shadow detection. Note that a significant portion of each shadow is detected as foreground. This overlapping allows our detection to be more robust by associating shadows with objects.

Figure 5 shows some foreground detection results on daytime and night-time images. The night-time scenes contain both areas where visible spectrum information is very useful (such as the streetlamp illuminated road in the centre right part of the image) and where it is not available (path areas). Our algorithm detects the passing car using primarily

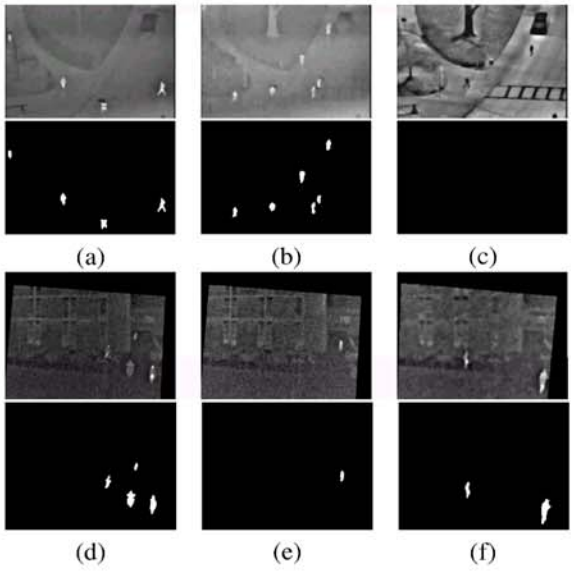

Figure 3: Pedestrian Detection Results: (a)-(c) Images from the 2005 OTCBVS Benchmark Dataset Collection. (d)-(f) Frames from our video data

visual information and detects the two people using thermal information.

In figures 6 and 7 we compare the effectiveness of our algorithm with and without the initial pedestrian detection and the Γ likelihood measure. The graphs were obtained using a ground-truth obtained by manually marking valid and invalid regions in every tenth frame. Without our improvements, false positives increase quickly as pedestrians move and leave ghost objects behind. No foreground pixels are missed by our model, until frame 400 when a person in a window is accidentally put into the background, as thermal radiation does not travel effectively through glass. Our pedestrian detection usually overestimates the foreground pixels, but quickly stabilises, as shown in figure 6 where the false positives drop quickly as non-pedestrians are put into the background model.

5. Summary and Conclusions

In this paper, we presented our approach to robust background modelling in the visible and thermal infrared spectra. We described our pedestrian detection algorithm that allows our model to initialise in the presence of foreground objects. We introduced our likelihood map, which accumulates evidence from size, edge and thermal features, suggesting that some foreground objects should be added to the background model, allowing our model to cope with ghosts and erroneous regions caused by lighting changes.

Visible spectrum background modelling has some inherent weaknesses, in that it relies on reflected light, therefore has difficulty in scenarios with uncontrolled lighting. Thermal infrared video, not only provides a complementary modality to the visible spectrum, but also provides the thermal features which can aid in the creation of the background model and verify the extracted foreground regions.

Figure 4: Shadow Detection: Visible spectrum frame, Infrared spectrum frame, detected foreground objects and detected shadows

Interestingly, while shadows do not occur generally in thermal video, a similar feature occurs in infrared that a person, or any hot object, is surrounded by a dark 'aura' due to the *chopper* within the infrared camera. Some interesting future work will involve detecting and eliminating this 'aura' as is currently done with shadows in the visible spectrum. Also, as our background model initialisation is quite effective, even with foreground objects in the scene, it may be beneficial to reinitialise the model at certain points during the system's operation. Having the ability to recognise when a large background error has been made, such as if the camera was moved, could trigger a re-initialisation of the model. Another additional improvement will be to use other features, such as motion, to update the background likelihood map.

Acknowledgments

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.

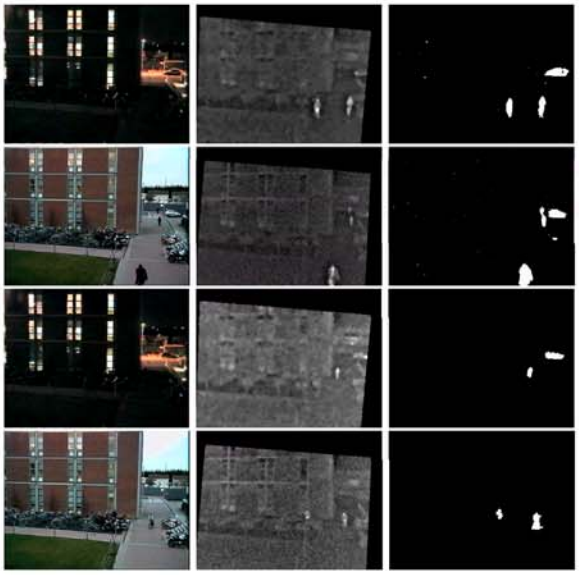

Figure 5: Foreground Detection

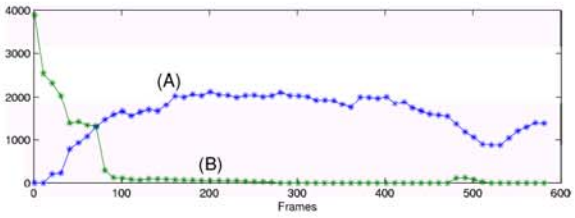

Figure 6: False positive foreground pixels obtained by (A) Background Model without pedestrian detection and likelihood accumulation (B) Our Model

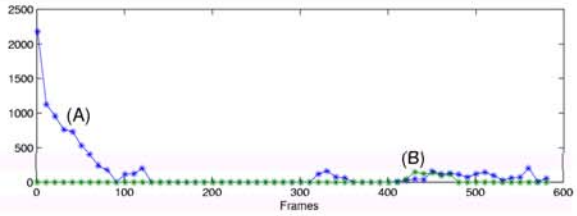

Figure 7: Missed foreground pixels obtained by (A) Background Model without pedestrian detection and likelihood accumulation (B) Our Model

References

[1] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *Proceedings of the 6th European Conference on Computer Vision*, 2000.

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

[3] T. Matsuyama, T. Ohya, and H. Habe. Background subtraction for non-stationary scenes. In *Proc. 4th Asian Conference on Computer Vision*, pages 662–667, 2000.

[4] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati. Detecting moving objects, ghosts and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, Oct 2003.

[5] 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.

[6] A. M. McIvor. Background subtraction techniques. In *Image and Vision Computing, Hamilton, New Zealand*, Nov 2000.

[7] A. Vetro, T. Haga, K. Sumi, and H. Sun. Object-based coding for long-term archive of surveillance video. In *IEEE International Conference on Multimedia and Expo (ICME)*, July 2003.

[8] S.-S. 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.

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

[10] 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.

[11] 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.

[12] A. Bovik. *Handbook of Image and Video Processing*. Academic Press, 2000. 621.367/BOV DCU Library Code.

[13] O. Javed, K. Shafique, and M. Shah. A hierarchical approach to robust background subtraction using color and gradient information. In *Workshop on Motion and Video Computing*, pages 22–27, Dec 2002.