TRACKING IN CCTV VIDEO USING HUMAN VISUAL SYSTEM INSPIRED ALGORITHM

M. Sugrue, E. R. Davies

Royal Holloway, University of London, UK.

{m.sugrue, e.r.davies}@rhul.ac.uk

Keywords: Motion, tracking, camera shake, HVS.

Abstract

Most recent motion tracking systems rely on subtraction of a background model from the current frame to leave moving objects segmented. Frame-to-frame correspondence is then used to match segmented regions and track objects. A great number of difficulties are involved in building and maintaining an accurate background model in the face of unpredictable lighting and weather conditions. Further, correspondence methods are not robust in matching non-rigid targets such as pedestrians. Here, a motion-tracking algorithm based on form–motion interaction similar to that used in the Human Visual System is proposed. This method avoids the difficulties of background modelling and tracks moving targets primarily based on their motion, rather than their appearance. This method is demonstrated in several CCTV pedestrian tracking scenarios.

1. Introduction

The recent proliferation of CCTV security systems (currently 3 million cameras in the UK alone) necessitates development of smarter software to take over the work done by human operators.

The most widely adopted approach relies on first, the production of a background model and then the subtraction of that model from the current frame to leave segmented the moving objects in the scene [3]. Segmented objects may then be tracked using frame-to-frame correspondence.

Background models are constructed from individual pixel statistical models, called pixel processes [4]. Taking past values of the pixel over a temporal window, the central tendency is used as the scene background value. If the pixel value differs from that pixel's central tendency by more than a threshold, the pixel is assumed to be part of an object that has moved in front of the background. Thus moving objects may be segmented.

The choice of the temporal window size as well as the choice of the central tendency statistic (temporal mean, temporal median or temporal mode filtering), depends on knowledge of the lighting and scene characteristics. Mean can give acceptable results in indoor environments with constant and diffuse lighting [10] whereas outdoor 'open-world' scenes require the greater robustness to lighting fluctuations offered by median or mode [4].

The Gaussian Mixture Model approach has become popular [7] as it has the advantages of allowing multimodal pixel processes, and thus a quick recovery time, and also a wide temporal window for robustness.

Background modelling schemes face two inherent and antagonistic problems, called the Stationary Background problem and the Transient Background problem [4]. First, the background model must reflect the stationary part of the scene to allow accurate segmentation of moving objects, required by frame-to-frame correspondence based tracking. This problem requires a low difference threshold in the subtraction stage and a wide temporal window so that slowly moving objects do not merge with the background. Second the background model must update to appearance changes in the scene, such as changed lighting conditions, requiring narrower temporal windows. No compromise gives perfect results, with a common failure being partial segmentation when the 'object depth' (object overlap in previous frames, equal to length/speed) is greater than half the temporal window. Similarly, a tracked object that then stops, such as some abandoned luggage, will merge with the background, vanishing from the tracking stage.

Numerous higher level algorithms have been proposed to 'reintegrate' partly segmented objects [12, 3] and deal with other difficulties of this methodology [7]. See for example Gavrila [5], who has had some success detecting pedestrians from moving vehicles using chamfer matching; and Siebel and Maybank [11] (whose human tracking work rests heavily on the early formative methods of Baumberg and Hogg [2]).

Tracking rigid targets using background modelling and frame-to-frame correspondence, such as in traffic monitoring applications, has been quite successful. Tracking pedestrians, however, is far more difficult due to the fact that the pedestrian changes shape and appearance while walking. Further, any inaccuracies in the segmentation step snowball into greater difficulties at this stage. Several systems have been recently proposed which combine background modelling, subtraction and multiple-blob particle filters [9].

1.1. Tracking in Human Visual System

A different computational paradigm is represented in the Human Visual System (HVS). The HVS has a remarkable ability to robustly track moving objects amid extreme motion clutter and adverse illumination conditions. While many important mysteries remain, years of neurological research have begun to reveal this system's secrets. A simplified overview of tracking in the HVS follows.

In the brain, the visual input stream divides quite early into two parallel processing streams: the dorsal pathway for motion information, and the ventral pathway for form information. In the dorsal pathway, sets of neurons are selectively responsive to horizontal and vertical motion of edges [6]. Other neurons are selective to certain speeds of motion. The form or ventral channel recognises stationary forms using a hierarchy of neural detectors that process form features of increasing complexity. The two channels interact at several levels, including pattern matching in tracking; however these interactions are not yet clearly understood [8].

When tracking a moving object the system can operate both in a predominantly feed forward mode, as in pre-attentive search, and the important top-down mechanism used in attentive tracking. Studies have shown that robustness to motion clutter in the HVS depends on attention.

Several computational models of these neurological structures have been recently developed [8, 6] with the aim of solving outstanding mysteries, namely the nature of the form–motion interactions.

The work in this paper focuses, conversely, on the potential practical uses for such models in CCTV tracking systems.

1.2. CCTV System Requirements

Smart CCTV systems require detection of events and human behaviour with robustness on par with a human operator. The UK Home Office's Police Scientific Development Branch is currently developing a Video Test Imagery Library (VITAL) to provide standards for the smart CCTV industry and researchers. VITAL has identified four application scenarios [1] for which new, more robust tracking and detection software is required. This paper presents results for "Detecting intrusions into sterile perimeter zones" and "Detecting when bags are abandoned." VITAL states its key requirements as robust outdoor performance under all weather and lighting conditions, along with a low false alarm rate due to shadows, clouds, animals, etc.

2. Motivation

The best video processing equipment available, capable of near perfect robustness and image understanding, is of course the Human Visual System (HVS). In recent years, neurologists, chiefly using fMRI, have been begun to peek into this black box and have revealed some of the startling ingenuity, as well as complexity, of the hardware inside. It is the motivation of this project to investigate whether these HVS methods might be useful in video applications.

The background modelling and subtraction methodology has become the standard model for motion segmentation in video. However, this methodology has a great number of inherent problems and provides a weak foundation for the later stages of tracking using frame-to-frame correspondence schemes. The main problem is that due to inaccurate background maintenance in the face of unpredictable lighting conditions, the segmented image always contains poorer information than the original image, building up difficulties for later stages. Further, neurology tells us the Background/Subtraction methodology does not have any known counterpart in the HVS. The motivation of this research is to investigate biology's solutions to the motion-tracking problem and develop an alternative to the background model.

3. Method

This CCTV system design aims to detect and track moving objects while avoiding difficulties and failings of the common background subtraction methodology. The system design replicates the operation of some parts of the HVS. The input video stream is split into a motion channel and a form channel. Target identification and tracking uses a form–motion interaction model similar to that found in the HVS. This is to avoid the problems involved in tracking non-rigid objects using frame-to-frame correspondence.

3.1 Computation of Motion Channel

The motion channel is calculated directly from the image stream using a simple volumetric method. Incoming frames are stacked to produce a 3-d array of pixels, of dimensions x, y, t. A 3×3 Sobel edge detector operator is used in the x, t and y, t planes. The result is a data array indicating location and speed of edge motion, in horizontal and vertical planes. This data is similar to that used by the HVS dorsal pathway.

This method gives a result similar to motion thresholding. Regions of the image with moving edges are emphasised, while regions without change are de-emphasised. Figure 2(a) shows the difference histogram for two consecutive frames in which a person is moving. A strong peak due to general image noise exists at greylevel 5, while other peaks are due to motion within the video. Figure 2(b) shows the histogram of the image result from the motion channel using the volumetric Sobel. It can be seen that the image has been split into stationary and moving regions. Figure 2(c) shows form and motion channel output.

3.2 Tracking Algorithm

The motion channel algorithm, a volumetric method described above, produces motion blobs at the locations of moving edges. Blobs are doubly defined as being above (1) a size threshold (8-connected) of 4 pixels and (2) an *activity–density* threshold. During operation, the tracking system remains idle until a motion blob above this double threshold

appears. The *activity–density* threshold is calculated from equation (1). The equation integrates the horizontal and vertical motion levels and normalises it for the pixel area, to produce the average motion per pixel. This threshold is necessary to eliminate scattered and low-level motion noise. The values of these thresholds are not critical and the system works well over a wide variation of threshold values. (For equation 1, the threshold must simply be between the two peaks of Fig. 2(b)) The threshold values used here have been shown to work well by experiment.

$$\frac{\sum_{Pixels}\sqrt{(HorizMotion)^2+(VertMotion)^2}}{NumberOfPixelsInBlob} > 100 \qquad (1)$$

All motion blobs above this threshold are then matched with the image at that location in the form channel (see schematic Figure 1). Where these motion blobs coincide significantly with the previous position of a tracked object, the form at the new location and the form of the object in the previous frame are compared. If they match the tracker is updated. If two or more separate regions of motion are detected where only one existed previously, the system will assume the object has split, as in the case of an individual separating from a crowd or a person dropping something (Figs. 3, 4).

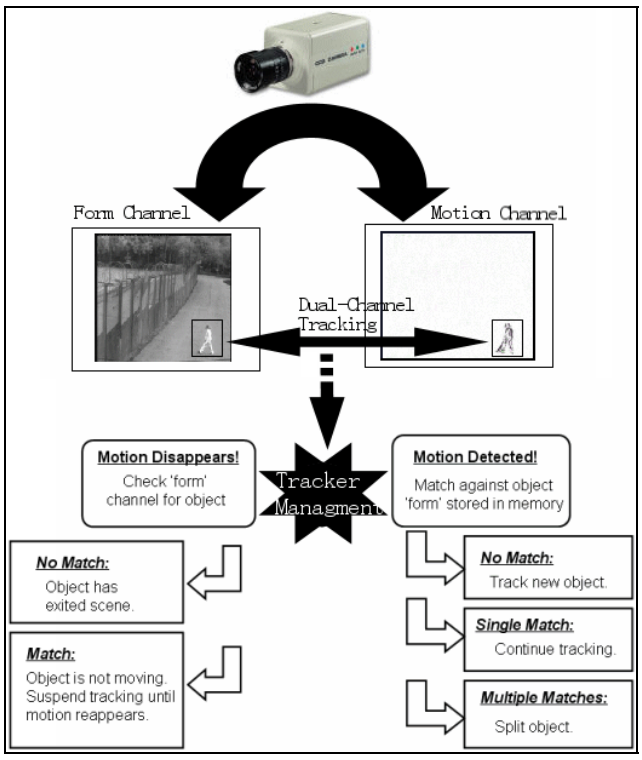
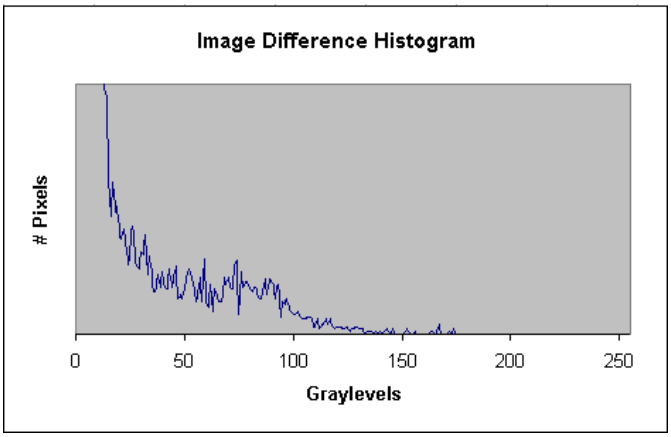
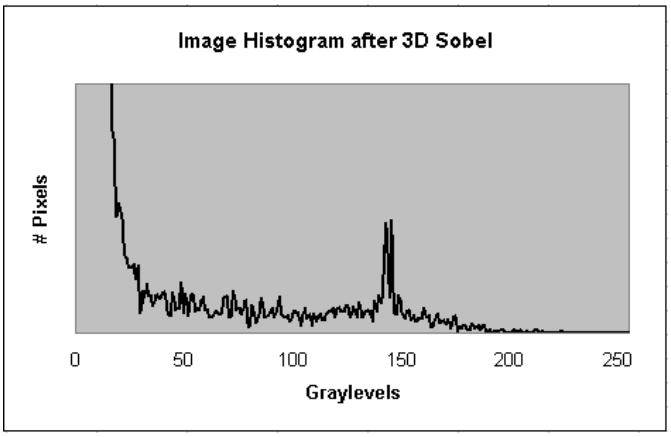


Fig. 1: Schematic diagram and algorithm flowchart of Form/Motion dual channel tracker system.

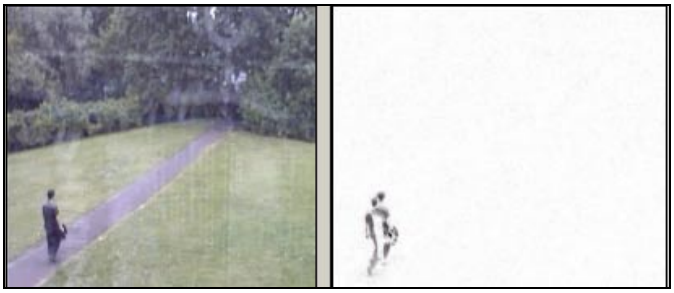
Traditional trackers based on search-and-match algorithms are hampered by the change in shape and appearance as the target person moves. The proposed system compensates for this using the motion channel to provide supplementary information.



(a)



(b)



(c)

Fig. 2: Image Difference Histogram and System Output. (a) Peak at zero is due to non-changing pixels. Other peaks are due to many kinds of pixel change, noise and lighting changes. (b) Peak at 145 represents edge motion; peak near zero is due to non-moving pixels. Compared with (a), all pixel change not due to edge motion has been shifted to the zero peak. (c) Left image shows frame with a moving pedestrian, right image shows output of motion channel. (a) and (b) are cropped on vertical axis for clarity.

The system deals with objects that come to rest as follows. If no motion blob exists in the motion channel at the location of a previously moving object, the tracker looks to the form channel to decide whether the object has stopped or has left the image. If the form of the object is still present, but has no motion, the object is defined as 'stale', and is monitored until it begins to move. If the aim is to detect suspect abandoned luggage an operator may be alerted after a certain time has lapsed.

The system works equally well for greyscale or colour video, and at any frame rate so long as an object is moving sufficiently slowly that there is an overlap from one frame to the next. No a priori knowledge of any structure within the image is assumed. A bootstrap time of only three frames is required before tracking begins.

Table 1 shows an algorithmic comparison between the common background/subtraction method and the proposed system.

Background/Subtraction Tracking	**New Form–Motion Tracking**
Bootstrap: Complex. Depends on background modelling method and temporal window. *Process steps per frame:* 1. Update background model using new frame: calculate per-pixel background value using central tendency of temporal window. 2. Subtract background model from current frame. Segment moving objects. 3. Use frame-to-frame correspondence to match segments based on appearance.	*Bootstrap:* 3 frames. *Process steps per frame:* 1. Calculate motion channel using volumetric Sobel detector. 2. Match motion 'blobs' to previous positions using size and shape of blob. In case of ambiguity, use 'form' channel match as well.

Table 1: Algorithm comparison between Background/Subtraction and Form–Motion methods of people tracking.

4. Results

The system was tested using both colour and greyscale video (between 10 and 25 frames per second (fps)) of pedestrians in outdoor scenes. Scenarios tested include intruder detection and tracking, abandoned luggage detection, multiple pedestrian tracking and tracking while the camera is shaking. The reason for taking four rather different scenarios was to demonstrate generality and to give a fair test to the tracking system that we have developed: it will be clear that the system is not geared merely to a single narrowly defined type of scene. Moreover, it is able to handle scenes that are of current interest to important bodies such as the Home Office.

The Figure 3 sequence shows output for tracking for a single intruder in VITAL's 'Sterile Zone' scenario. The target is tracked as soon as it enters the frame and until it leaves. In a practical system, an intruder alarm could be sounded when the target comes too close to the fence.

The Figure 4 sequence shows how the system tracks a walking person. When part of the target separates (in this case the tracked person drops their coat) the item is tracked separately. If it comes to rest (i.e. it is only present in the form channel), the system monitors it until it begins to move again. This function of the system may be used to monitor abandoned luggage and to provide information on its origin.

Figure 5 shows multi-target tracking of pedestrians. The system performs well, tracking all six moving pedestrians present in the 180-second test video. A weakness of the current implementation can be seen from the Frame 1. The 'squiggle' in the centre of frame one of Figure 5 shows the point where two pedestrians stood chatting for a few moments before separating. As the motion channel highlights only moving regions and the centroid of these regions are tracked, the 'squiggle' track represents their gesture movements. When these pedestrians themselves began to walk away their whole bodies were detected by the motion channel and tracked.

Also visible in Figure 5 is a failing due to the use of a global threshold level. The two figures at the top right were not immediately detected as they were walking slowly directly towards the camera. Their motion was detected in the Motion Channel but fell below the global threshold used to eliminate noise from the tracker. The system could be improved by replacing the global motion threshold with a threshold that can be locally set.

Figure 6 shows results of tracking a single target when the CCTV camera is shaking. Camera shake is very common in practical CCTV systems, as they are often positioned on high poles. Shake can cause significant difficulties when background modelling and subtraction is used in tracking. The system tracked the pedestrian correctly initially when the camera shake was at about 10% (measured as the fraction of the frame 'cropped'). When the shaking was increased briefly the system lost the track but found it again by Frame 4. The gap in the track is visible in Frame 4.

5. Discussion

Figures 3 and 4 show robust tracking under varied lighting conditions and video quality. Figure 4 also demonstrates abandoned luggage detection. These represent important requirements of the Home Office VITAL project. Further, these results were obtained without the need for complex and error-prone background modelling or the search-and-match algorithms of traditional trackers.

Also unlike traditional trackers, this scheme does not use a Kalman filter or other predictive device (as used for example by Siebel and Maybank [11]). A Kalman filter was employed

during the early development of this system but was found to be unnecessary. Because objects are tracked primarily in the motion channel, their direction of motion can be directly computed. However, no provision has yet been made for consistent object labelling following splitting or occlusions. This extension will form part of the future work.

This system also shows itself to be quite robust to camera shake. Camera shake is an important problem for practical CCTV systems, as they are often positioned on high poles. Figure 6 shows the results of tracking a single person under conditions of mild shaking (~10%). The basic system as described above is still successful in tracking the human target while using no special compensation functions to correct for camera shake. To improve the system's robustness to camera shake, compensation can be added using a simple function that computes frame motion from the motion channel data.

One weakness of the system is that an object may sometimes be incorrectly detected twice. As only the moving edges of an object are present in the motion channel, an object that moves exactly parallel to the frame axis may have its leading and trailing edges separated, and thus be detected as two separate objects. The HVS faces an identical problem at this stage. Neurons that detect moving edges do so locally. The solution is a higher level of neurons that integrate local edge information to the object level. A future extension to this system will be a higher-level integration step to join incorrectly divided objects.

6. Conclusions

This paper has described a new form–motion channel approach to image sequence analysis and object tracking that is inspired by the known functions of the HVS. This has a number of advantages over the usual background/subtraction approach to motion segmentation and tracking and should be seen as an alternative approach to background modelling.

The system's use of motion edge detection and tracking avoids many of the difficulties and complexities inherent in the background modelling and subtraction approach. Also, with a temporal window of only three frames, the system achieves a practical balance between recovery time and robustness to change.

Traditional trackers based on search-and-match algorithms are hampered by the change in shape and appearance as the target person moves. The proposed system compensates for this using the motion channel to provide supplementary information.

While approaches may be inspired by the HVS, in machine vision applications it is important than any system that is adopted be effective and efficient, not merely that it adhere to some idealised schema that may turn out to be suboptimal: our findings are that the latter eventuality is far from being the case with our system.

Future work will include extending the system to deal with object occlusion, the addition of an object level integration step to avoid the problems of multiple detections of objects, replacing the global activity threshold level with a locally variable level, and the addition of global frame motion compensation to improve robustness to camera shake.

Acknowledgements

This research has been funded by Research Councils UK under Basic Technology Grant GR/R87642/02.

The authors would also like to thank the Home Office Police Scientific Development Branch for providing some early release VITAL test videos.

References

[1] D. Baker, "Imaging to Fight Crime", Police Scientific Development Branch, http://www.homeoffice.gov.uk/docs2/vital_solution.html (website accessed 26/11/2004).

[2] A. Baumberg, D. Hogg, "An adaptive eigenshape model", Proc. British Machine Vision Assoc. Conf. (Sept.), pp. 87–96 (1995).

[3] A. Bevilacqua, "A novel background initialization method in visual surveillance", IAPR Workshop on Machine Vision Appl., Nara, Japan, pp. 614–617 (2002).

[4] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, "Detecting moving objects, ghosts, and shadows in video streams", IEEE Trans. Pattern Anal. Machine Intell., **25**, no. 10, pp. 1337–1342 (2003).

[5] D. Gavrila, "Pedestrian detection from a moving vehicle", Proc. European Conf. on Computer Vision, D. Vernon (ed.), Dublin, Ireland (June), pp. 37–49 (2000).

[6] M.A. Giese, T. Poggio, "Neural mechanisms for the recognition of biological movements", Nature Neuroscience, **4**, pp. 179–192 (2003).

[7] E. Grimson, C. Stauffer, "Learning patterns of activity using real-time tracking", IEEE Trans. Pattern Anal. Machine Intell., **22**, no. 8, pp. 747–757 (2000).

[8] S. Grossberg, E. Mingolla, L. Viswanathan "Neural dynamics of motion integration and segmentation within and across apertures", Vision Research, **41,** pp. 2521–2553 (2001).

[9] M. Issard, J. MacCormick, "BraMBLe: A Bayesian multiple-blob tracker", Proc. Int. Conf. Computer Vision, **2**, pp. 34–41 (2001).

[10] R. Mattone, A. Glaeser, B. Bumann "Advanced Video Surveillance" in *Multimedia Video-Based Surveillance*, Eds. G. L. Foresti, P. Mahonen, C. S. Regazzoni, Int. Ser. Eng. Comp. Sci. Pub: Kluwer, (2000).

[11] N.T. Siebel, S.J. Maybank, "Fusion of multiple tracking algorithms for robust people tracking". In A. Heyden, G. Sparr, M. Nielsen, P. Johansen (eds.) Proc. 7th European Conf. on Comp. Vision (ECCV), vol. IV, pp. 373–387 (2002).

[12] K. Toyama, J. Krumm, B. Brumitt and B. Meyer, "Wallflower: principles and practice of background maintenance", Proc. 7th Int. Conf. on Computer Vision, Corfu, Greece. Vol. 2, pp. 255–261 (1999).

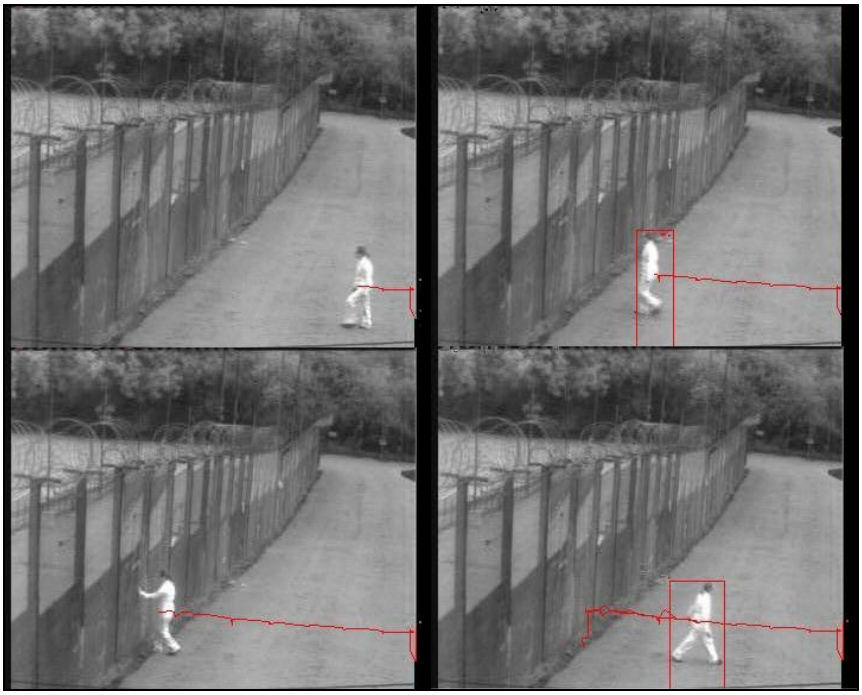


Fig. 3: Single intruder tracking. Figure shows four sample frames from a 30 second greyscale video sequence at 25fps. The red line represents the path of the target centroid, as calculated by this system. *(Video courtesy of Home Office)*

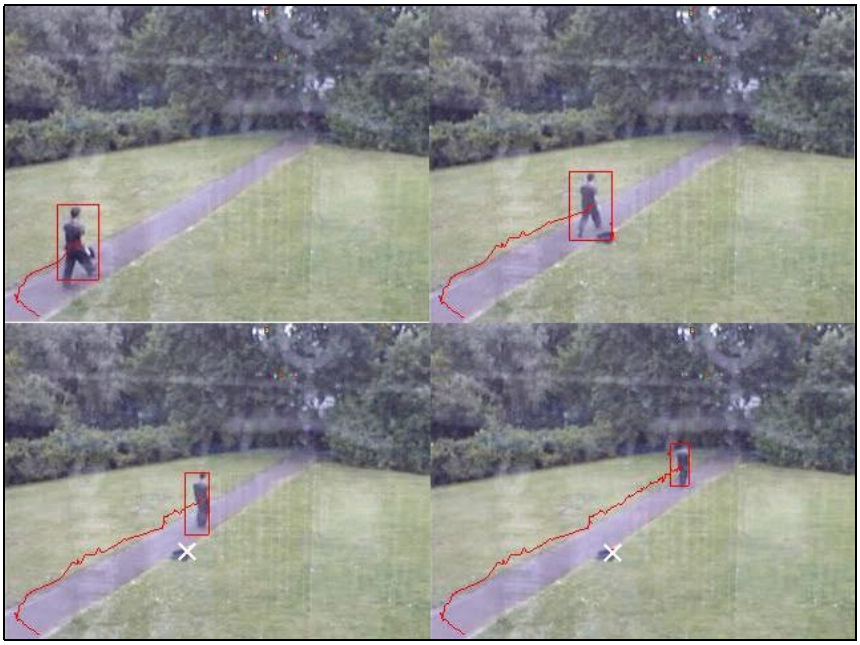


Fig. 4: Single target discards an item. Figure shows four sample frames from a 60 second colour video sequence at 10 fps. Red line and box show the path and bounding box of the moving target, white shows the location of the stationary target. The discarded item is individually tracked once it separates from the main 'blob'.

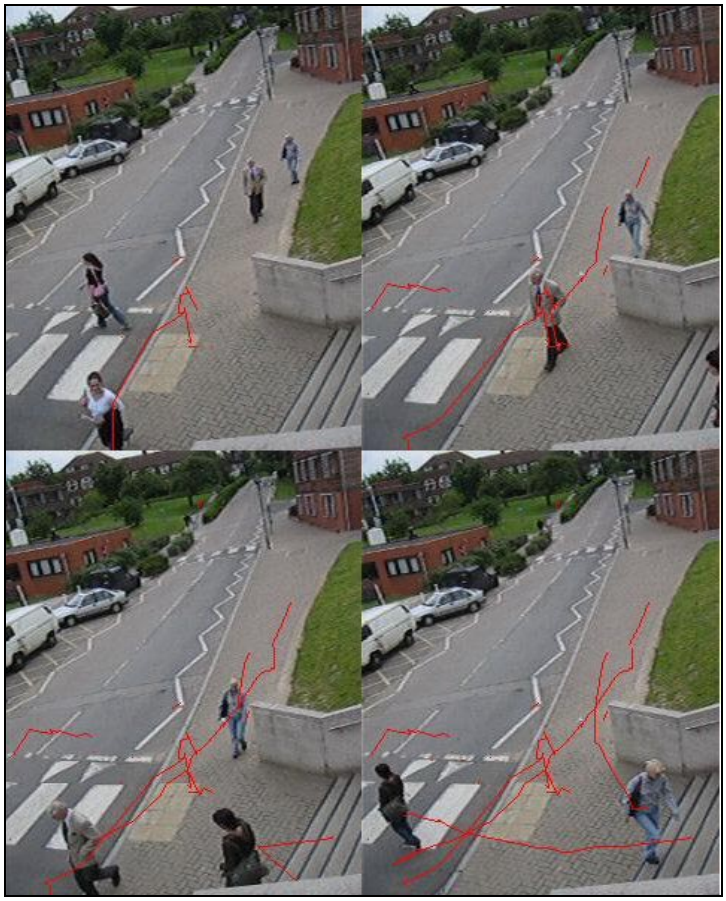


Fig. 5: Tracking Multiple Targets. Figure shows four sample frames from a 180 second video at 25fps of multiple pedestrian tracking in a street scene.

Fig. 6: Tracking during camera shake. Figure shows results of tracking a single target when the CCTV camera is shaking.