HUMAN DETECTION AND TRACKING THROUGH TEMPORAL FEATURE RECOGNITION

Fraser K. Coutts, Stephen Marshall, and Paul Murray

Department of Electronic and Electrical Engineering, University of Strathclyde
Royal College Building, 204 George Street, G1 1XW, Glasgow, Scotland
phone: + 44 (0) 141 548 2205, fax: +44 (0) 141 552 2487, email: fraser.coutts.2013@uni.strath.ac.uk
web: www.strath.ac.uk/eee/research/cesip/

ABSTRACT

The ability to accurately track objects of interest – particularly humans – is of great importance in the fields of security and surveillance. In such scenarios, the application of accurate, automated human tracking offers benefits over manual supervision. In this paper, recent efforts made to investigate the improvement of automated human detection and tracking techniques through the recognition of person-specific time-varying signatures in thermal video are detailed. A robust human detection algorithm is developed to aid the initialisation stage of a state-of-the-art existing tracking algorithm. In addition, coupled with the spatial tracking methods present in this algorithm, the inclusion of temporal signature recognition in the tracking process is shown to improve human tracking results.

Index Terms— Automated human tracking, thermal video, temporal characteristic recognition

1. INTRODUCTION

Automatically detecting and tracking humans is of interest in many application areas, particularly in security and surveillance operations. Thermal imagery is particularly useful for this purpose as it can operate under any level of target illumination, and it delivers high contrast between humans and their surroundings due to temperature differences. Furthermore, it can allow the effects of different weather conditions to be mitigated.

A number of tracking algorithms capable of following the path of a moving person have been presented in literature. Four state-of-the-art, open source, tracking algorithms have been identified as being applicable to thermal data [1-4]. These trackers are known as MILTrack [1], Incremental Visual Tracking (IVT) [2], Tracking-Learning-Detection (TLD) [3], and Real-Time Compressive Tracking (RTCT) [4].

During operation, MILTrack, IVT, and TLD all rely on the incremental improvement of an appearance model of a target to account for appearance or lighting changes. Thus, if a target's shape is altered during the tracking process, each algorithm is able to learn the new form of the object of interest while operational.

RTCT incorporates a very sparse measurement matrix to efficiently extract the features for the updating of an appearance model. Compression arises during the application of a random matrix in its feature extraction process, which effectively reduces the dimensionality – and therefore the size – of the data extracted. It is shown in [4] that RTCT generally outperforms other state-of-the-art trackers – including MILTrack and TLD – when applied to a set of benchmark testing data.

While useful for single person tracking, the above algorithms were found to be susceptible to target loss when encountering occlusion and situations involving multiple humans in thermal video. Furthermore, an automated tracker initialisation stage was desirable, as current trackers typically require an initial bounding box to be manually entered.

We have observed that certain fluctuations in the intensity of pixels within thermal image sequences can be indicative of the presence of humans and their associated movements. In this paper, we present a robust method for detecting and tracking individuals within thermal video sequences using both spatial and temporal features of a targeted human. The proposed technique builds upon the RTCT tracking algorithm, due to its demonstrated reliability. A standalone human detection algorithm based on Support Vector Machine (SVM) [5] classification of temporal features automates the initialisation procedure of the RTCT tracking algorithm, and modifications are made to this algorithm to accommodate simultaneous spatial and temporal feature evaluation during tracking. While other methods of spatio-temporal tracking and detection have been investigated for use in thermal imagery [6-9], it is believed that the temporal signatures used in this paper introduce a new concept.

A specially recorded dataset containing thermal videos of multiple humans was captured for testing purposes.

Results obtained show the successful automation of the tracker's initialisation, and that the inclusion of temporal feature recognition offers improvements in human tracking results, suggesting that such techniques may be viable in future tracking efforts.

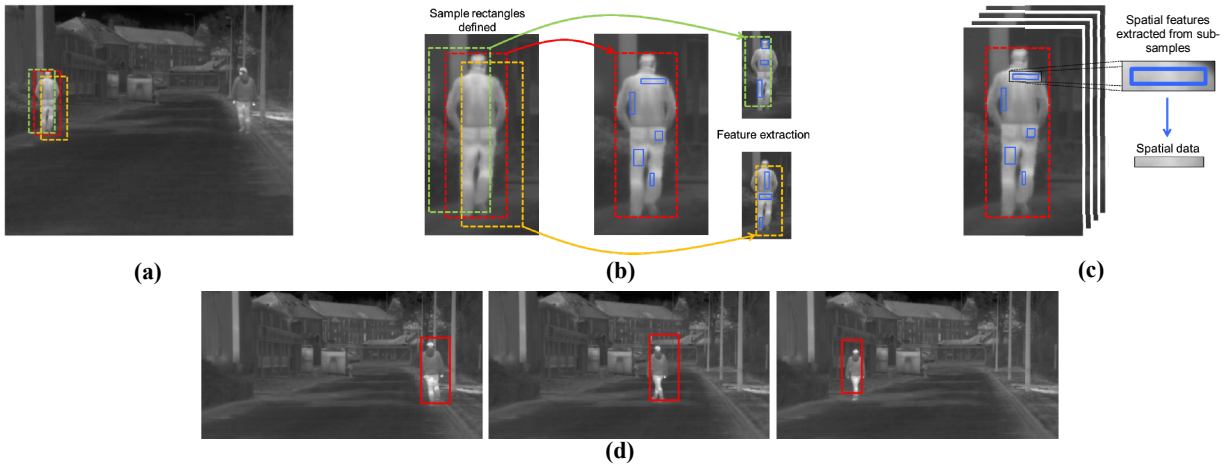


Fig 1. Real-Time Compressive Tracking. **(a)** Sample rectangles located around the existing bounding box. **(b)** Sub-samples taken from within sample rectangles. **(c)** Spatial feature extraction from defined sub-samples. **(d)** Stages of human tracking using the RTCT algorithm. Human moves across the field of view from right to left.

2. DESCRIPTION OF DATA

The input data used for the testing of the existing and developed human tracking algorithm was in the form of 12, 16-bit thermal videos; all had a resolution of 640x512 pixels, and an effective frame rate of 25 frames per second. These videos were tailored to depict different scenarios in an attempt to emulate possible situations in the real world and to give images of humans at various ranges.

Simple two-dimensional matrices sufficed to store individual frames of a video sequence, but for temporal analysis, multiple frames were stored simultaneously in three-dimensional matrices. In this case, the third dimension contained time-varying thermal data and was of a size determined by the length of the sequence being analysed.

3. REAL-TIME COMPRESSIVE TRACKING

While a detailed description of the operation of RTCT is provided in [4], a brief summary of the main components relevant to the modifications described in this paper is provided here to minimise the need for cross-referencing.

The initialisation stage of the RTCT algorithm requires a bounding box – which surrounds the object to be tracked in subsequent frames – to be defined by a pre-existing text file. This file contains the location and size of the box to be used. Once initialised, the algorithm begins tracking the object within the specified box.

During tracking, the RTCT algorithm determines a number of sample rectangles defined around the location of the current bounding box. When transitioning between frames, the sample rectangle containing features that most closely match previously observed characteristics of the object being tracked is chosen as the new bounding box. The tracking algorithm then updates the characteristics of the tracked object by analysing sample rectangles near to the current bounding box, and learns features of the surrounding background by recording the contents of rectangles placed further afield. Figure 1(a) illustrates the placement of sample rectangles around an original bounding box.

For each sample rectangle, a feature extraction template is applied and a number of sub-samples obtained. This process is shown in by Figure 1(b), where the smaller internal rectangles are representative of the sub-samples taken. Following this, spatial data can be extracted from the sub-samples within each rectangle and used to form a feature list that defines the spatial characteristics of the overall sample. Figure 1(c) demonstrates this procedure. Newly acquired spatial samples are compared with previous positive (human) and negative (non-human) samples and allocated a confidence rating, which indicates how well they match the target of interest.

Figure 1(d) shows the RTCT algorithm successfully tracking a human target walking across the view of the camera. The algorithm's need for an initial bounding box requires prior knowledge of the target's location or a manual input stage. Thus, an algorithm to automatically detect the presence of humans and present the tracker with an initial bounding box is desirable. Furthermore, while the RTCT algorithm worked well for simple human tracking applications, it was prone to failure when the target human was occluded by another in a scenario henceforth identified as a 'cross-over'. In Section 5, we demonstrate that by including time varying thermal data in the list of features computed by the RTCT tracker, this issue can be overcome.

4. HUMAN DETECTION ALGORITHM

When analysing pixel intensities over time, pixels containing humans typically display greater variance than those containing the relatively static background, as Figure 2(a) demonstrates. It was therefore considered viable that a human detection algorithm could be developed that specifically identifies any time-varying signatures similar to those of a human.

Principal Component Analysis (PCA) [10] is a mathema-

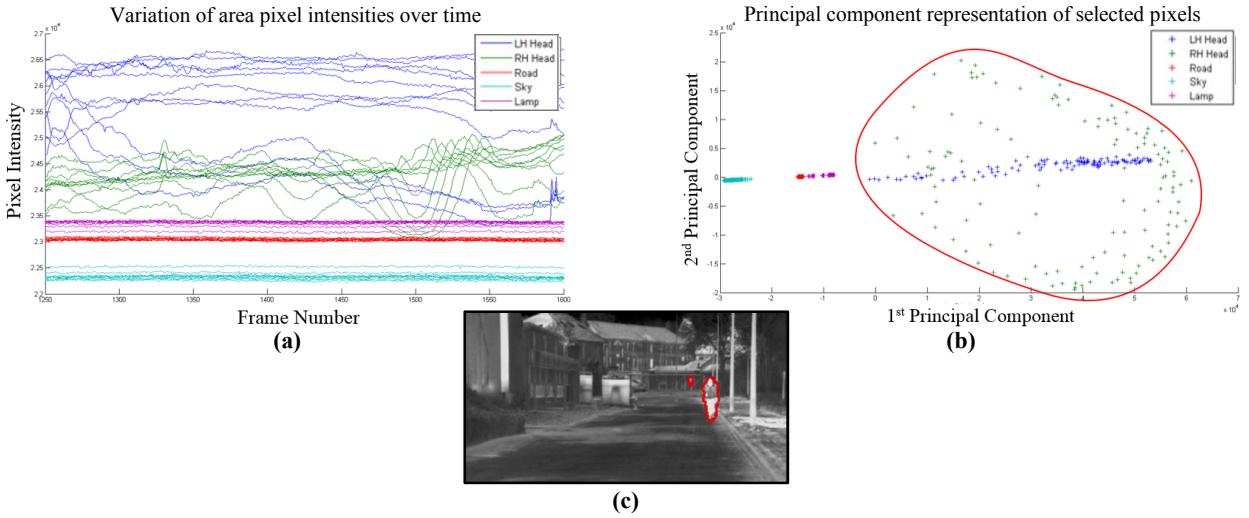


Fig 2. Human detection based on temporal signature of humans. **(a)** Pixel intensity fluctuations of selected areas against time. Higher intensity and variance fluctuations belong to pixels containing humans (LH Head, RH Head); lower intensity and variance fluctuations belong to pixels containing background objects (Road, Sky, Lamp). **(b)** First principal component versus the second for each pixel's fluctuation. Pixels containing humans are highlighted. **(c)** Output frame identifying presence of humans within red outlines.

tical data processing technique that can be employed to reduce the dimensionality – and therefore the size – of a data set prior to further processing. Here, PCA can be used on time-varying pixel intensities extracted from video sequences to reduce the dimensionality of the variance of the pixel intensities over time. This process enables the separation of dissimilar video features – such as humans and static background objects – based on their respective variances. Figure 2(b) highlights the separation observed when plotting the first two principal components against each other following the application of PCA to the time-varying pixel intensities of Figure 2(a).

SVMs are a form of supervised learning binary classifier that can be trained to recognise the characteristics of positive and negative features – in this case, human and non-human traits – and subsequently be deployed on unseen data sets where they can detect the presence of these features [5]. Their initial training stage uses data of known class and chosen parameters to determine the orientation and placement of support vectors, which define the boundary that separates the two classes in the given spatial dimensions.

For human detection in thermal video, PCA can be used to reduce the dimensionality of datasets prior to their input to the SVM for training or classification purposes. As the classification of each pixel within a frame is a time consuming process, steps must be taken to reduce the number of pixels to be classified.

One way to reduce the number of pixels to be processed is to apply motion detection to eliminate pixels exhibiting insufficient movement from the classification process. This results in a large portion of the background being ignored. For a sequence of frames, the mean intensities of each pixel can be calculated. These average values can then be subtracted from corresponding pixel intensities in the initial frame. Subsequent thresholding of the resulting image can then eliminate pixels with low variance.

Blob detection can also be used to limit the pixels to be analysed to those contained within regions of similar intensity – it is assumed that any humans present in a thermal video will give rise to such regions. The Maximally Stable Extremal Regions (MSER) [11] method of blob detection was used for this application. MSER blob detection operates by incrementing through an image's intensity profile by a set threshold delta; changes in size of different regions within the image as this occurs are calculated and used to judge if each region is stable or not. Stable regions are then classified as blobs within the image. The maximum possible area of blobs can be restricted to the maximum expected size of humans within the image sequences used. An SVM is then applied to the blobs in the image to determine if they are human in origin.

When applying this algorithm to a video sequence containing two humans – one running towards the camera and another far in the distance – the algorithm gave the result shown in Figure 2(c). It can be seen that the method has correctly placed red outlines around the locations of the two humans. For the initialisation of the RTCT algorithm, the intensity or number of pixels contained within any outlined areas can be used to specify the characteristics of the human to be tracked. For example, if tracking of a nearby human is preferential, the area containing the largest number of pixels can be automatically selected by the algorithm as the initial bounding box for the tracker.

5. MODIFIED REAL-TIME COMPRESSIVE TRACKING

We propose in this paper a Modified RTCT (MRTCT) algorithm which uses extracted temporal and spatial features to track humans in thermal video. The extraction of these fea-

tures is enabled by the modification of the existing feature extraction functionality of the RTCT algorithm. Once extracted, both spatial and time varying thermal signatures can be used to form a feature list that defines the characteristics of a sample.

For comparison of temporal features in MRTCT, additional functionality must be implemented. The technique of cross-correlation [12] is considered viable for these purposes. Of particular interest is the cross-correlation coefficient, which is a measure of the strength of relationship between two normalised datasets. Datasets that progress in an identical fashion incur a cross-correlation coefficient of 1, while datasets that progress in an opposite fashion result in a value of -1. A cross-correlation coefficient of zero implies that there is no relationship between the tested datasets.

Through determination of the cross-correlation coefficients linking newly taken samples with previously recorded positive and negative samples, the direct comparison of time-varying signatures is possible. Following cross-correlation of temporal features in successive frames, a confidence rating can again be associated with each potential bounding box to quantify its temporal similarity to the object of interest.

With confidence ratings for each sample available from both spatial and temporal classifiers, a simple cost function enables the generation of a new parameter. The peak value of this parameter identifies the sample rectangle that possesses the spatial and temporal characteristics most similar to that of the desired target.

A decay function to weight previous temporal characteristics reduces the overall impact that any erratic training samples have, while allowing a temporal profile of the target to be built upon over time. Within this function, a learning rate parameter determines the speed at which the tracker adapts to changes in the object of interest's temporal features.

The inability of the RTCT algorithm – and therefore the MRTCT algorithm – to shrink an object's bounding box as it moves further away was thought to influence its operation at longer distances. While the bounding box may contain only the tracked person upon initialisation, a large portion of the box will contain the background as this person moves away from the camera.

To shrink or expand the bounding box as required, the aforementioned sample rectangles generated during the tracking process and their associated confidence ratings – which indicate the likelihood of a rectangle containing the object of interest – can be used. For example, if multiple rectangles carry a high confidence rating while being in close proximity, it is likely that a new bounding box can be generated using a combination of their areas.

6. EXPERIMENTAL RESULTS

Initial bounding boxes generated using the human detection algorithm of Section 4 were successfully used within the RTCT and MRTCT trackers. Figure 3(a) demonstrates the acquisition of an initial bounding box through human detection, and Figure 3(b) and (c) show its recognition within the tracking algorithm and the beginning of the tracking process, respectively.

The scenario shown in Figure 4 shows the failure of the RTCT algorithm to track the correct person following a cross-over. It was hypothesised that this result could be improved upon through application of the proposed spatial and temporal data fusion method for tracking. During testing of this hypothesis, the initial bounding box – which was automatically generated using the human detection algorithm of Section 4 – and parameters relevant to the spatial feature recognition were kept constant.

Early results – obtained when bounding box shrinking functionality was not present – indicated that in some cross-overs, the inclusion of temporal data improved correct target selection, but the problem of incorrect selection was still present.

The addition of functionality to shrink the bounding box improved results considerably for both the RTCT and MRTCT algorithms; however, MRTCT consistently performed well, and in some cases better than the unmodified tracker. Figure 5 demonstrates the ability of both trackers to track the same person in a sequence containing a large number of cross-overs at various ranges. The individual frames of Figure 5(a) and (b) allow direct comparison of the trackers' performance at stages within this sequence. Both trackers achieve a similar level of accuracy for the majority of the sequence, but in the bottom frames of Figure 5 it can be seen that MRTCT has outperformed RTCT; at this point, MRTCT has tracked the correct target while RTCT has failed.

The Centre Location Error (CLE) – as used in [4] – is defined as the distance between the central locations of the tracked target and the manually labelled ground truth for a sequence. Use of the CLE enabled a quantitative comparison of the accuracy of MRTCT and RTCT, with the two algorithms achieving average CLEs of 11.6 and 25.1 pixels, respectively, for the sequence of Figure 5.

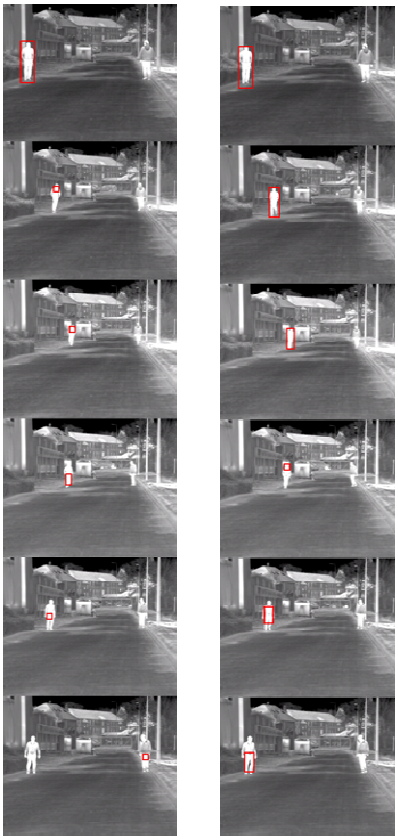
CONCLUSIONS

This paper presents the successful automation of a state-of-the-art tracking algorithm's initialisation stage through the application of a human detection algorithm based on temporal feature recognition. The automation of a tracker's initialisation stage in this way is of great importance if the tracker is to operate autonomously.

Furthermore, the modification of the RTCT algorithm, known as MRTCT – which accommodates simultaneous spatial and temporal feature evaluation during tracking – has given promising results that suggest there may be a case for future research in this area.

(a) **(b)** **(c)**

Fig 3. Automated bounding box initialisation. **(a)** Human detection through classification. **(b)** Bounding box initialises tracking. **(c)** Tracking operational.



(a) **(b)**

Fig 5. Tracking of human using **(a)** RTCT and **(b)** MRTCT.

(a) **(b)** **(c)**

Fig 4. Tracking of human using RTCT tracker. **(a)** Initial bounding box located on person. **(b)** Correct person tracked initially. **(c)** Incorrect person tracked following cross-over.

ACKNOWLEDGEMENTS

The authors would like to thank Stephen McGeoch, Barry Connor, and Christopher Dickson of Thales UK for the use of their high performance thermal imaging equipment and data, and for their valuable guidance during the development of this project.

REFERENCES

[1] B. Babenko, M.H. Yang, and S. Belongie, "Robust Object Tracking with Online Multiple Instance Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619-1632, August 2011.

[2] D.A. Ross, J. Lim, R.S. Lin, and M.H. Yang, "Incremental Learning for Robust Visual Tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125-141, May 2008.

[3] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-Learning-Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1409-1422, July 2012.

[4] K. Zhang, L. Zhang, and M.H. Yang, "Real-Time Compressive Tracking," in *ECCV 2012*, Florence, 2012, pp. 864-877.

[5] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[6] J. Han and B. Bhanu, "Human Activity Recognition in Thermal Infrared Imagery," in *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, San Diego, 2005, pp. 17-24.

[7] A. Leykin, Y. Ran, and R. Hammoud, "Thermal-Visible Video Fusion for Moving Target Tracking and Pedestrian Classification," in *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, Minneapolis, 2007, pp. 1-8.

[8] C. Bregler, "Learning and Recognizing Human Dynamics in Video Sequences," in *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, San Juan, 1997, pp. 568-574.

[9] C.N. Padole and L.A. Alexandre, "Wigner Distribution based Motion Tracking of Human Beings using Thermal Imaging," in *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, San Francisco, 2010, pp. 9-14.

[10] I. Jolliffe, *Principal Component Analysis*, John Wiley & Sons, Ltd, 2005.

[11] J. Matas, O. Chuma, and T. Pajdlaa, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, September 2004

[12] A.G. Asueroa, A. Sayagoa, and A.G. González, "The Correlation Coefficient: An Overview," *Critical Reviews in Analytical Chemistry*, vol. 36, no. 1, pp. 41-59, 2006.