The Influence of Multi-Sensor Video Fusion on Object Tracking Using a Particle Filter

L. Mihaylova, A. Loza, S. G. Nikolov, J. J. Lewis,
E.-F. Canga, J. Li, T. Dixon, C. N. Canagarajah and D. R. Bull

mila.mihaylova@lancaster.ac.uk, {artur.loza, stavri.nikolov}@bristol.ac.uk

Abstract: This paper investigates how the object tracking performance is affected by the fusion quality of videos from visible (VIZ) and infrared (IR) surveillance cameras, as compared to tracking in single modality videos. The videos have been fused using the simple averaging, and various multiresolution techniques. Tracking has been accomplished by means of a particle filter using colour and edge cues. The highest tracking accuracy has been obtained in IR sequences, whereas the VIZ video was affected by many artifacts and showed the worst tracking performance. Among the fused videos, the complex wavelet and the averaging techniques, offered the best tracking performance, comparable to that of IR. Thus, of all the methods investigated, the fused videos, containing complementary contextual information from both single modality input videos, are the best source for further analysis by a human observer or a computer program.

1 Introduction

Recently there has been an increased interest in object tracking in video sequences supplied by a single camera or a network of cameras [FMS+05, HBC+05, PVB04, BK05, YSS03]. Reliable tracking methods are of crucial importance in many surveillance systems as they enable human operators to remotely monitor activity across areas of interest, enhance situation awareness and help the surveillance analyst with the decision-making process. Multiple-sensor systems can provide surveillance coverage across a wide area and under different conditions. The complementarity of the information supplied by different sensors is valuable for solving different kinds of problems, such as detection, recognition, tracking and situation assessment, to name but a few, and ought to be fused in order to serve more efficiently these purposes. *Data fusion* comprises groups of methods for merging data from multiple sensors. In this paper, we investigate how the fusion of video sequences from visible (VIZ) and infrared (IR) cameras influences the process of object tracking. We compare the tracking results from the fused sequences with the results obtained from separate IR and VIZ videos.

The remaining part of this paper is organised as follows. Section 2 presents the methodology used to assess the particle filter tracking performance in multi-sensor real video data. Conclusions and future plans are given in Section 3.

2 The Assessment of Object Tracking in Fused Videos

2.1 Video Sequences, Fusion and Tracking Techniques Used

This paper investigates the influence of the fusion process on the Particle Filter (PF) tracking developed in [BMCB06, BMCB05]. Informative image features, such as colour, edge,

texture and motion cues can be used in combination, or adaptively chosen, for a reliable performance. In this particular implementation [BMCB06], we use colour and edge cues, with the Bhattacharyya distance as a measure of distance between the target and each current region [CRM03, Bha43]. However, other distances, such as the distance based on the structural similarity measure [LMCD06], can capture spatial information from the image and can provide better performance than the Bhattacharyya distance and the histogram-based approach, especially in presence of illumination variations.

The experiments presented in this paper have been performed with real-world data collected from one VIZ and one IR camera. The testing sequences, taken from the Eden Project multisensor dataset of short range surveillance videos [LNL$^+$06] (available at *www.scanpaths.org*), are very challenging for automatic object tracking due to the following: (a) the colour of the moving object being very similar to the background (camouflaged moving targets); (b) the environment containing lush and dense vegetation; (c) frequent obscuration of the targets; and (d) the very low signal-to-noise ratio of the VIZ videos. The videos were fused using the following methods: Averaging Technique (AVE), Contrast Pyramids (CP), Discrete Wavelet Transforms (DWT) and Dual-Tree Complex Wavelet Transforms (DT-CWT) (see [DLN$^+$06] for more details on these techniques).

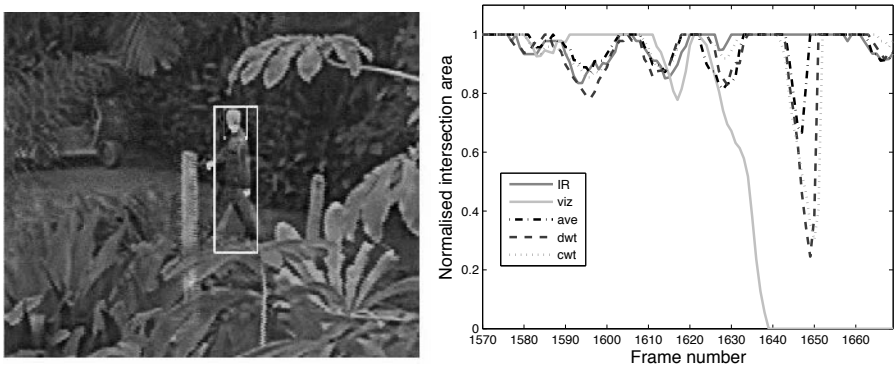

Figure 1: a) Left: ground truth rectangle (the big box, surrounding the whole person) [DLN$^+$06] and PF reference rectangle (the smaller box). b) Right: plot of the normalised intersection area S for the tested trackers.

2.2 Performance Evaluation Measure

In order to assess the influence of video fusion on the accuracy of the object tracking, video sequences with pre-drawn 'target maps' [DLN$^+$06] have been used. These are rectangular boxes drawn around the target (a walking human figure) on each frame (see Figure 1a). The estimate of the target position, and corresponding estimated rectangular box, is obtained from the tracker (the small rectangle in Figure 1a). The target map (ground truth rectangle) and the estimated rectangle vary in size from frame to frame. In order to measure the performance of the tracker, the normalised intersection area between the two rectangles has been used. The rationale for such a measure is that a larger intersection/overlap between the estimated area (B) and the ground truth rectangle (A), indicates better tracking. This normalised intersection area S is calculated as $S = (A \cap B)/B$. The measure varies between 0 (no intersection between the rectangles) and 1 (estimated rectangle fully within ground truth rectangle).

2.3 Experimental Results with Real Data

The PF tracker with 300 particles has been run with the same non-optimised default parameters in all experiments, in order to be able to compare object tracking performance with different video sources. A rectangle used to manually initialise the PF tracker has been selected to be significantly smaller than the ground truth rectangle, so that only the core features of the object of the interest (the head and the upper part of the body) are included, while the influence of the nonstationary background is minimised.

The measure S has been calculated for each frame and all the methods investigated and its plots are shown in Figure 1b. In order to obtain the global characteristics of the tracker performance, the difference between the actual and 'ideal' tracker has been calculated: $\sum_k (1 - S_k)$, where k is a frame index. According to this validation the most accurate tracking results are obtained with the IR video (3.6). Other methods, such as fused AVE (4.8), DT-CWT (6.8), DWT (7.9) performed comparably, whereas single modality VIZ (38.7) suffered from frequent loss of the target. Accordingly, Figure 1b shows also that the best overlap between the two rectangles is obtained from the IR video, then the AVE, the DT-CWT and finally the DWT. The worst result is obtained with VIZ. The colour-edge based algorithm failed to track the object in the VIZ videos due to the fact that it is almost the same colour as the background.

Figures 2-5 show the results from tracking in VIZ, IR, AVE, and DT-CWT videos, respectively, from one of the Tropical Forest sequences used in this study. The results from the CP are not given because of the many artifacts introduced by the fusion, rendering the tracker unreliable. The results from the DWT are not shown because of their similarity to the DT-CWT.

Figure 2: A tracker working with the VIZ video: frames 1570, 1625 and 1670. The object is lost after the full occlusion.

Figure 3: A tracker working with the IR video: frames 1570, 1625 and 1670. The object is lost after the full occlusion, but recovered after that.

Figure 4: A tracker working with AVE fused video sequences: frames 1570, 1625 and 1670. A reliable performance is observed.

Figure 5: A tracker working with DT-CWT fused video sequences: frames 1570, 1625 and 1670. A reliable performance is observed.

It was observed that the best trackers have been able to keep tracking the object despite its severe occlusions. Such an example is shown in Figures 2-5, where the person is completely hidden by a tree. Since the head features are well distinguishable from the surrounding environment, the tracker succeeded in recovering the person. The tracker may still loose the object in the presence of long full occlusions (5-10 frames or more) or when the target region did not correspond completely to the initially chosen reference region. This could be solved by adding a detection scheme to the tracker and updating the reference region.

The obtained results are similar to findings of a psycho-visual study, conducted independently, regarding experiments with human eye tracking [DLN+06] over the same sequences of data. In this case the AVE also produced some of the best results.

3 Conclusions

The results reported in this paper show that IR videos provide the highest object tracking accuracy. It is argued however, that the fused videos, although containing some artefacts that degrade the tracking performance to a small degree, are a good alternative to tracking in single modality data. This is due to inclusion of the complementary and contextual information from all input sources, that make it more suitable for further analysis by either a human observer or a computer program. Of all the investigated fusion methods, the AVE and DT-CWT techniques have been found to outperform the DWT, CP, and VIZ tracking.

Among the issues that need further investigation is automatic re-detection of the object after a lengthy full loss and comparison with human eye movement results of a target pursuit task in the same sequences. Tracking with multiple cameras of the same fused modality will also be investigated in the future. In such case, the computational complexity of the algorithm working with multiple views has to be considered and the problem of dealing with the redundant information has to be addressed.

Acknowledgements. The authors are grateful to the UK MOD Data and Information Fusion Defence Technology Centre funding this research.

References

[Bha43] A. Bhattacharayya. On a Measure of Divergence Between Two Statistical Populations Defined by Their Probability Distributions. *Bull. Calc. Math. Soc.*, 35:99–110, 1943.

[BK05] E. Blasch and B. Kahler. Multiresolution EO/IR Target Tracking and Identification. In *Proc. of the International Conf. on Information Fusion*, pages 275–282. ISIF, 2005.

[BMCB05] P. Brasnett, L. Mihaylova, N. Canagarajah, and D. Bull. Particle Filtering with Multiple Cues for Object Tracking in Video Sequences. In *Proc. of SPIE's 17th Annual Symp. on Electronic Imaging, Science and Technology, V. 5685*, pages 430–441, 2005.

[BMCB06] P. Brasnett, L. Mihaylova, N. Canagarajah, and D. Bull. Sequential Monte Carlo Tracking by Fusing Multiple Cues in Video Sequences. *Image and Vision Computing, in print*, 2006.

[CRM03] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564– 577, May 2003.

[DLN+06] T.D. Dixon, J. Li, J.M. Noyes, T. Troscianko, S.G. Nikolov, J. Lewis, E.-F. Canga, D.R. Bull, and C.N. Canagarajah. Scanpath Analysis of Fused Multi-Sensor Images with Luminance Change: A Pilot Study. In *9th International Conf. on Information Fusion*. ISIF, Italy, 2006.

[FMS+05] G. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, and T. Ellis. Active Video-based Surveillance System: the Low-level Image and Video Processing Techniques Needed for Implementation. *IEEE Signal Processing Magazine*, 22(2):25– 37, March 2005.

[HBC+05] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti. Smart Video Surveillance: Exploring the Concept of Multiscale Spatiotemporal Tracking. *IEEE Signal Processing Magazine*, 22(2):38–51, March 2005.

[LMCD06] A. Loza, L. Mihaylova, N. Canagarajah, and Bull D. Structural Similarity Measure for Object Tracking in Video Sequences. In *Proc. of the 9-th International Conf. on Information Fusion*, Florence, Italy, 10-13 July 2006.

[LNL+06] J. Lewis, S. Nikolov, A. Loza, E.-F. Canga, N. Cvejic, J. Li, A. Cardinali, N. Canagarajah, D. Bull, T. Riley, D. Hickman, and M. I. Smith. The Eden Project multi-sensor data set. Technical Report TR-UoB-WS-Eden-Project-Data-Set, University of Bristol UK and Waterfall Solutions Ltd, April 2006. http://imagefusion.org/.

[PVB04] P. Pérez, J. Vermaak, and A. Blake. Data Fusion for Tracking with Particles. *Proceedings of the IEEE*, 92(3):495–513, March 2004.

[YSS03] A. Yilmaz, K. Shafique, and M. Shah. Target Tracking in Airborne Forward Looking Infrared Imagery. *Image and Vision Computing*, 21(7):623–635, 2003.