RGB-D and Thermal Sensor Fusion:
Application in Person Tracking

Ignacio Rocco Spremolla[1,2*], Michel Antunes[1], Djamila Aouada[1], and Björn Ottersten[1]
[1]*Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg*
[2]*MVA Master Programme, ENS Cachan, Université Paris-Saclay, France*
iroccosp@ens-cachan.fr, {michel.antunes, djamila.aouada, bjorn.ottersten}@uni.lu

Keywords: Sensor Fusion; RGB-D; Thermal Sensing; Person Tracking

Abstract: Many systems combine RGB cameras with other sensor modalities for fusing visual data with complementary environmental information in order to achieve improved sensing capabilities. This article explores the possibility of fusing a commodity RGB-D camera and a thermal sensor. We show that using traditional methods, it is possible to accurately calibrate the complete system and register the three RGB-D-T data sources. We propose a simple person tracking algorithm based on particle filters, and show how to combine the mapped pixel information from the RGB-D-T data. Furthermore, we use depth information to adaptively scale the tracked target area when radial displacements from the camera occur. Experimental results provide evidence that this allows for a significant tracking performance improvement in situations with large radial displacements, when compared to using only a tracker based on RGB or RGB-T data.

1 Introduction

There are many applications that make simultaneous use of visual data and other sensing modalities. In the past few years, extensive research has been carried out for fusing RGB and Depth sensors (RGB-D). A non-exhaustive list of examples where this type of multi-modal systems are employed include human pose estimation (Shotton et al., 2011), action recognition (Vemulapalli et al., 2014), simultaneous localization and mapping (Endres et al., 2012), and people tracking (Luber et al., 2011). Combining RGB and thermal sensors (RGB-T) has been investigated to a lesser extent, and mainly used for robust person tracking (Stolkin et al., 2012; Kumar et al., 2014).

Currently, approaches that combine the three sensor modalities discussed previously are being investigated (Mogelmose et al., 2013; Vidas et al., 2013; Nakagawa et al., 2014; Matsumoto et al., 2015; Susperregi et al., 2013). Each modality supports a particular capability: depth sensors provide real-time and robust 3D scene structure information; RGB cameras provide rich visual information; and thermal sensors allow to compute discriminative temperature signatures. In this paper, we pursue this line of work, and extend the RGB-T fusion scheme presented in (Talha and Stolkin, 2012) to RGB-D-T. We present a calibration pipeline (intrinsics and extrinsics) that allows to fuse the information acquired by the different sensor modalities, and propose a person tracking algorithm based on RGB-D-T data. Experimental results provide evidence that combining the three modalities for the purpose of people tracking using a traditional particle filter based framework is more robust when compared to using only a single or pairs of sensors.

1.1 Contributions

A common problem when fusing the data captured with multiple sensor modalities is to find the corresponding regions between the data. In this paper, we address this problem and present a pipeline for the intrinsic and extrinsic calibration of all the sensors (RGB, D and T). This allows to generate registered RGB, depth and thermal images, which have a one-to-one pixel correspondence. From the best of our knowledge, using such mapped data for the purpose of person tracking has not been addressed before. Moreover, this data registration process is not particular to the followed tracking approach, and could be applied as a pre-processing step for other tracking algorithms as well.

Furthermore, we present a simple scheme based on a particle filter for fusing the RGB-D-T data for the

*The totality of this work was performed while the author was an intern at SnT, University of Luxembourg.

purpose of person tracking. Compared to the previous work based on RGB-T data only (Talha and Stolkin, 2012), we use the additional depth information in two different ways: 1) in computing additional descriptors extracted from the depth data; and 2) in continuously adapting the tracked target size with an appropriate scaling factor (adaptive scaling), yielding increased robustness to radial motion.

1.2 Organization

This paper is organized as follows: in the next section, we briefly discuss the most relevant works that use RGB, RGB-D, RGB-T or RGB-D-T data for the purpose of object or people tracking. In Section 3, we present the experimental setup, the calibration procedure, and the process of image registration. Section 4 discusses the theory behind the proposed person tracking algorithm, and how we take advantage of the different sensor modalities. Finally, Section 5 presents the experimental results, where the tracking accuracy using RGB-D-T is evaluated with respect to using only RGB or RGB-T.

2 Related Work

This section briefly reviews the most relevant works that use RGB, RGB-D, RGB-T or RGB-D-T data for tracking purposes.

In the past, extensive work has been done in object tracking from RGB video sequences (Pérez et al., 2002; Nummiaro et al., 2002; Nummiaro et al., 2003). More recently, with the popularization of RGB-D sensors, objects or people trackers based RGB-D data have started to be studied. In practice, the use of depth sensors makes the process more robust against illumination changes at a lower computational cost, and the provided 3D structure information simplifies many tasks such as background subtraction and object segmentation. Choi et al. (Choi and Christensen, 2013) describe a system for detecting people from image and depth sensors on board of a robot, where detection algorithms using the two different sources of information are fused using a sampling based method. Choi and Christensen (Choi and Christensen, 2013) present a particle filtering approach for object pose tracking, where the likelihood of each particle is evaluated using features extracted from RGB and D data. Going one step further, Jafari et al. (Jafari et al., 2014) present a multi-person detection and tracking system suitable for mobile robots and head-worn cameras. The authors use an extended Kalman filter framework and use different types of algorithms for extracting relevant information from the multi-modal data (e.g. 3D point classification, visual odometry, RGB based sliding window pedestrian detection).

There are also multi-modal systems based on RGB-T data for tracking applications. Stolkin et al. (Stolkin et al., 2012) present a Bayesian fusion method for combining pixel information from thermal imaging and conventional colour cameras for tracking a moving target. Very recently, Kumar et al. (Kumar et al., 2014) integrate a low-resolution thermal sensor with an RGB camera into a single system. The basic idea is to apply an RGB tracker and use the thermal information for eliminating a variety of false detections. Talha and Stolin (Talha and Stolkin, 2012) employ a particle filter tracking approach that fuses both sources of data, and adaptively weights the different imaging modalities based on a new discriminability cue.

Only very recently, RGB-D-T based systems have started to be used. Some applications include person re-identification (Mogelmose et al., 2013), and 3D temperature visualization (Vidas et al., 2013; Nakagawa et al., 2014; Matsumoto et al., 2015). From the best of our knowledge, there is only one work that fuses RGB-D-T data for people tracking and following (Susperregi et al., 2013). The algorithm is based on a particle filter for merging the information provided by the different sensors, which includes also a Laser Rangefinder. The overall pipeline includes special detectors, such as an emergency-vest and a leg detector. Our aim is different in the sense that we want to fuse the multi-modal RGB-D-T data following a low-level strategy, not requiring application specific high-level reasoning, such as special environment or target based detectors.

Our work extends the previously discussed framework (Talha and Stolkin, 2012) for the case of RGB-D-T data, where the additional depth information provides additional cues that, as evaluated experimentally in Section 5, shows to be very effective for the purpose of person detection and tracking.

3 Experimental Setup

This section briefly introduces the experimental setup we used to acquire and map RGB-D-T data.

The hardware used for this research was a Microsoft Kinect v2 RGB-D sensor, and a FLIR A655sc Thermal camera. The two sensors were positioned side-by-side using a rigid support, as depicted in Fig. 1. Using this setup, the following image data is acquired:

- RGB or colour image (C): 1920×1080 pixels at

30 fps[2],

- Depth/IR image (D): 512×424 pixels at 30 fps,
- Thermal image (T): 640×480 pixels at 50 fps.

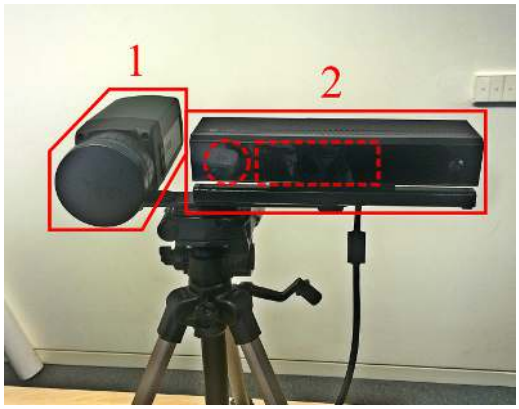


Figure 1: The setup: 1. FLIR A655sc thermal camera, 2. Microsoft Kinect v2 RGB-D sensor (dashed circle: RGB camera, dashed rectangle: ToF depth/IR sensor)

In order to fuse the data captured by each sensor, two steps are required: sensor calibration and image registration. These steps will be explained in the following sections.

3.1 System Calibration

System calibration is required in order to find the transformation of 3D points to image points in each camera. This transformation involves two steps, the determination of the pin-hole camera model parameter matrix K (intrinsic calibration) and the estimation of the relative sensor poses $[\mathsf{R}, \mathbf{t}]$(extrinsic calibration).

Before we move on, it is important to mention that the depth stream from the Kinect v2 comes from a time-of-flight camera that produces an additional IR stream from the amplitude information. As the two streams come from the same sensor, we call it the Depth/IR sensor. Although we do not use the IR stream for tracking purposes, we do use it for calibrating the sensor. This IR stream does not contain any thermal information, and should not to be confused with the stream coming from the thermal camera.

The cameras are calibrated using the well-known toolbox of Bouguet (Bouguet, 2004). The following steps are performed:

- Intrinsic calibration of the RGB camera using a checkerboard pattern, obtaining its intrinsic matrix K_C,

[2]fps: frame per second.

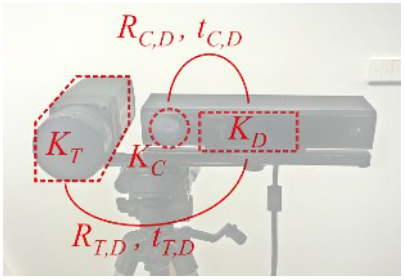


Figure 2: The intrinsic and extrinsic calibration parameters

- Intrinsic calibration of the Depth/IR sensor using a checkerboard pattern, obtaining its intrinsic matrix K_D,
- Intrinsic calibration of the Thermal camera using a disjoint squares pattern, obtaining its intrinsic matrix K_T,
- Computation of the relative pose of the RGB camera with respect to the Depth/IR sensor using a checkerboard pattern, obtaining the transformations $[\mathsf{R}_{C,D}, \mathbf{t}_{C,D}]$,
- Computation of the relative pose of the Thermal camera with respect to the Depth/IR sensor using a disjoint squares pattern, obtaining the transformations $[\mathsf{R}_{T,D}, \mathbf{t}_{T,D}]$.

The computed calibration parameters and transformations are illustrated in Figure 2. Note that the depth camera frame is used as the reference frame to extrinsically calibrate the RGB and the thermal cameras. In the calibration steps involving the RGB and/or the Depth/IR cameras, a standard paper checkerboard pattern was used, as it is visible in both modalities. However, for the calibration steps involving the thermal camera, a disjoint-squares pattern with cut-out squares and placed against a thermal backdrop was employed, as previously done in (Vidas et al., 2012).

3.2 Image Registration

In order to find the corresponding regions in the different image modalities, we perform a registration process on the raw images from the different sensors to generate registered C, T and D images having pixel-to-pixel correspondence. This is accomplished by computing a 3D point cloud from the depth image, and then projecting each 3D point to the RGB and Thermal images to assign to it a colour and thermal intensity. These points are then re-projected to the depth image plane, obtaining two new images: one with the corresponding RGB colour information, and the other with the corresponding thermal intensities. The final registered images, having pixel-to-pixel correspondences are shown in Figure 3

(a) Color image (RGB data) (b) Thermal image (T data) (c) Depth image (D data)

Figure 3: Registered images with pixel-to-pixel correspondence

The advantage of using this strategy is that the registered set of images simplifies the subsequent fusion step, and the pixel-to-pixel correspondence persists, even when the target has significant radial motion. It is important to note, however, that some pixels are marked as invalid in the final images (e.g. red points in Figure 3(a)), corresponding to pixels that are out-of-range or detected as being occluded in at least one of the cameras. These unassigned pixels are discarded when computing the descriptors for tracking.

4 RGB-D-T based Person Tracking

This section describes the tracking algorithm which we applied with the RGB-D-T system presented in the previous section, for the purpose of person tracking.

In order to assess the advantages of the RGB-D-T system, we employ a simple particle filter approach, and avoid using specific motion models and complex particle re-sampling strategies based on previous frames.

For each frame, the probability distribution of the target is estimated using a discrete set of N particles, where the impact of each particle i is appropriately weighted using w^i. Each particle i is defined by its centre position (p_x^i, p_y^i), and its foreground region width and height (p_w^i, p_h^i), respectively. It also encodes a local background region, defined between the foreground region, and an outer rectangular region of dimensions $(\lambda p_w, \lambda p_h)$, where λ is a user defined constant. A particle example is shown in Figure 4.

The estimated target state, or the tracker output for each frame, is computed by weighting the particle features:

$$p_\alpha = \sum_{i=1}^{N} p_\alpha^i w^i,$$

where p_α corresponds to any of the parameters p_x, p_y, p_w, p_h.

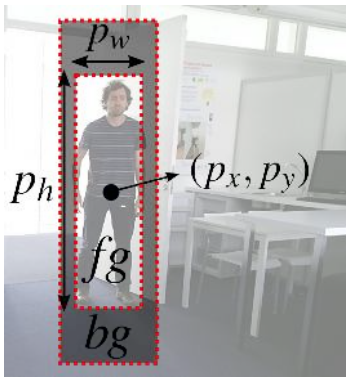


Figure 4: Each particle is defined by the parameters p_x, p_y, p_w, p_h of the foreground region (fg), and by a background region (bg).

4.1 Data Descriptors

For each particle and for each image modality, the foreground and background regions are used for computing appropriate descriptors. As described next, we use histogram based descriptors because they are fast to compute, and exhibit some invariance to rotations, partial occlusions and moderate non-rigid deformations (Nummiaro et al., 2002).

RGB based descriptor. For the colour modality, we convert the RGB image to a normalized colour-space which we will denote by rgb, where $r = R/(R+G+B)$, $g = G/(R+G+B)$, and $b = B/(R+G+B)$. The colour normalization discards the illumination information for achieving robustness to lighting changes. Since two components are enough for characterizing the normalized colour space ($r+g+b=1$), we compute a 2D histogram H_C using the pair (r,g).

D based descriptor. For the depth modality, we first compute a 3D normal vector for each data point by fitting a 3D plane to a pre-defined local neighbourhood. A 2D histogram H_D is then computed using

the corresponding polar angle θ and azimuthal angle φ information.

T based descriptor. For the thermal modality, we compute a 1D histogram H_T using directly the intensity values of the thermal image.

All the three descriptors are computed for the manually-selected target region in the first frame of the sequence, and defined as being the target model $\{H_C^t, H_D^t, H_T^t\}$. The size of the target region (t_w, t_h) and its mean depth t_d are also stored and later used for constructing appropriate particle hypotheses. Remark that the target model is not relearned along the sequence.

For each particle hypothesis, the three descriptors are computed using the corresponding foreground region, obtaining $\{H_C^{fg}, H_D^{fg}, H_T^{fg}\}$. Additionally, the same descriptors are also computed for the particle background, obtaining $\{H_C^{bg}, H_D^{bg}, H_T^{bg}\}$.

A usual measure for the comparison of two histograms $p = \{p^{(i)}\}_{i=1:n}$ and $q = \{q^{(i)}\}_{i=1:n}$ of n bins, is the Bhattacharyya similarity coefficient $B[p,q]$ (Nummiaro et al., 2003):

$$B[p,q] = \sum_{i=1}^{n} \sqrt{p^{(i)} q^{(i)}}$$

We employ it for the comparison of particle and target histograms. Each particle foreground descriptor is compared against the correspond target model descriptor, obtaining three coefficients $\{B_C^{fg,t}, B_D^{fg,t}, B_T^{fg,t}\}$ for the colour, depth and thermal data, respectively. By computing these coefficients, we are comparing the particle foreground descriptors with the target model, and are able to assess its similarity. Next, the same histogram comparison approach is used for analysing the particle background descriptors and the target model, obtaining three coefficients $\{B_C^{bg,t}, B_D^{bg,t}, B_T^{bg,t}\}$. These coefficients are used assessing how similar the background is to the target model, which is called the level of *camouflaging* of that particular particle.

4.2 Multi-modal Fusion

In order to fuse the information from the different modalities and determine the overall appropriateness of each particle, we extend the RGB-D based method presented in (Talha and Stolkin, 2012) to RGB-D-T data. The idea is to compute an enhanced Bhattacharyya coefficient B_f, which combines the parameters $B_C^{fg,t}$, $B_D^{fg,t}$ and $B_T^{fg,t}$ in a way such that less weight is given to modalities where *camouflaging* occurs. The computation of B_f is done as follows:

$$B_f = \alpha B_C^{fg,t} + \beta B_T^{fg,t} + \gamma B_D^{fg,t},$$

where

$$\alpha = \frac{B_T^{bg,t} + B_D^{bg,t}}{2(B_C^{bg,t} + B_T^{bg,t} + B_D^{bg,t})},$$

$$\beta = \frac{B_C^{bg,t} + B_D^{bg,t}}{2(B_C^{bg,t} + B_T^{bg,t} + B_D^{bg,t})},$$

$$\text{and} \quad \gamma = \frac{B_C^{bg,t} + B_T^{bg,t}}{2(B_C^{bg,t} + B_T^{bg,t} + B_D^{bg,t})}.$$

By doing so, we are essentially computing a weighted average of $B_T^{fg,t}$, $B_C^{fg,t}$ and $B_D^{fg,t}$ using adaptive weights that depend on the *camouflaging* of the complementary modalities. This ensures the desired effect of reducing the importance of modalities that could increase the uncertainty of the estimation approach.

Regarding the particle weight assignment, we first apply an exponential function on B_f, with the aim of stretching the range of values

$$\hat{w} = e^{-(1-B_f)/(2\sigma^2)},$$

where $\sigma = 0.2$ is an empirical constant. The final weight w^i of the ith particle is computed by normalizing $\hat{w}^i$ by the sum of the weights over the N particles:

$$w^i = \frac{\hat{w}^i}{\sum_{j=1}^{N} \hat{w}^j}.$$

4.3 Adaptive Target Scale

In many tracking applications, the tracked objects do not present significant radial distance changes with respect to the camera reference frame when compared to their lateral displacements. This produces the "size" of the tracked object to remain approximately constant along the tracking sequence.

In this work, we tried to address the problem of tracked objects at a short range from the camera, and which can present significant radial motion. In order to address this issue, we used the depth information to adjust the width and height of the particle window. This is achieved by extracting an image section from the depth image around each particle centre using the previous particle size (p_w, p_h). Then, a histogram is computed using the depth values from this region, and the particle depth centre p_d is estimated by determining the position of the mode of the histogram. Finally, the particle size

$$\begin{aligned} p_w &= t_w s, \\ p_h &= t_h s, \end{aligned} \tag{1}$$

is adjusted by scaling the target size (t_w, t_h) using a scale-factor s, based on the depth-ratio between the target size and depth

$$s = \frac{t_d}{p_d}. \quad (2)$$

5 Experimental Evaluation

For the experimental evaluation, we use three video sequences of a person moving in an indoor scene, which are composed of about 200 frames each, and were taken at 20fps. All the sequences include significant radial motion.

The first sequence corresponds to a person performing a diagonal motion, going from the far right to the near left of the scene. The trajectory described by the person can be observed in Figure 5(a). In the second sequence, the person moves closer and farther from the camera twice, and the returns to the original position. The trajectory is shown in Figure 5(b). Finally, in the third sequence, the person is initially positioned at a large distance from the camera setup (≈ 6.3m), and then performs a sequence of fast movements around the room (including jumps). This trajectory can be observed in Figure 5(c).

5.1 Evaluation

For the evaluation of the proposed tracking algorithm, ground-truth data was generated for the three video sequences. For this, a human operator manually selected a rectangular region covering the tracked person from head to feet and shoulder width, for one every five frames of each video. The parameters for the intermediate frames were then linearly interpolated.

Regarding the tracking quality, we decided to use a measure that considers the area of overlap between the ground-truth target region and the tracked region. For this, the Jaccard index was selected:

$$J(A,B) = \frac{A \cap B}{A \cup B}.$$

From the definition, we can see that the Jaccard index is bounded between 0 and 1, where 0 corresponds to completely disjoint regions (no overlap), and 1 corresponds to identical regions (perfect overlap). This measure is widely used in the literature, e.g. in (Everingham et al., 2010) for object category recognition and detection.

5.2 Results

For each video, three different sensor combinations were considered: RGB, RGB-T and RGB-D-T. Moreover, for each of these options, two runs were conducted: one using constant window size, and another one using adaptive window size scaling as described in Section 4.3.

For each case, the Jaccard index of the tracked region against the ground-truth region was computed. The results can be observed in Figures 6 and 7, while in Figure 8 the mean Jaccard index values are shown.

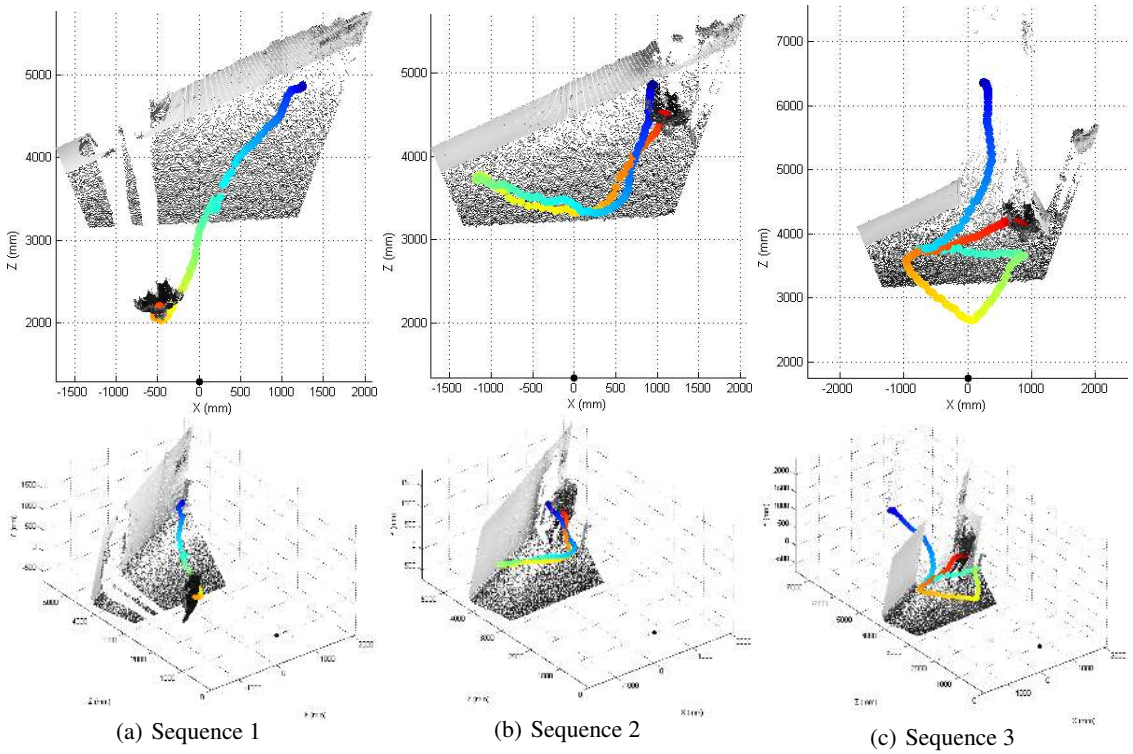
As expected, the accuracy is considerable low when using constant target sizes due to the radial motion of the person moving in the scene. The combinations RGB-T and RGB-D-T show superior performance when compared to the single RGB modality.

For the case where adaptive scaling is used, we can see that the three sensing combinations have considerably better performance in all the sequences. Note that this is only possible in a multi-modal framework where depth data is available. Furthermore, the descriptors based on RGB-D-T show slight overall accuracy improvements when compared to the other approaches.

6 Conclusions

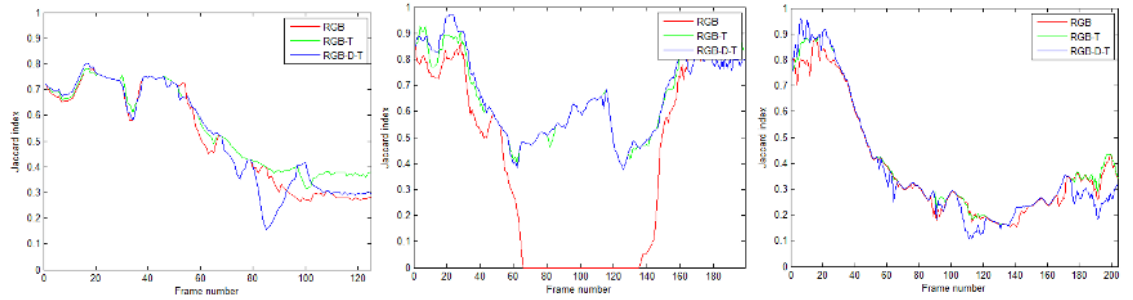
We have investigated the problem of fusing the data captured by a low-cost RGB-D camera with a thermal sensor. We showed how to completely calibrate the multi-modal system, and proposed a simple person tracking algorithm using mapped RGB-D-T data. By using the depth data to adaptively scale the target size, we proved that the tracker can resist to significant radial motions with good accuracy based on the Jaccard index. Moreover, we presented a simple way to extend the RGB-T tracker presented in (Talha and Stolkin, 2012) to RGB-D-T, by using a histogram of 3D normals as depth descriptor. Although the depth feature we used did not significantly improve the accuracy of the tracker in the tested video sequences, we believe it could improve its robustness in other more complicated sequences involving the interaction of several persons.

In this work, we modelled the target model using a single histogram for each data source. An interesting extension would be to use a multi-part model, and investigate how to efficiently compute histogram descriptors for each target using part specific fusion schemes. Finally, we the usage of a depth descriptor based on local shape information such as curvature distributions, instead of 3D normals, could add additional robustness to human deformations.



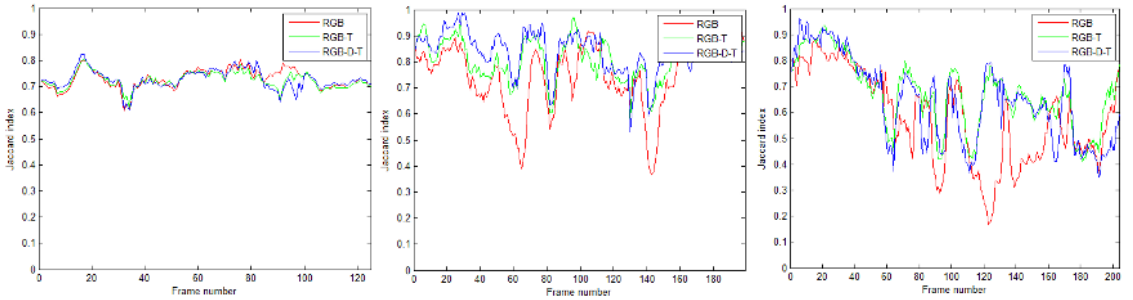
(a) Sequence 1 (b) Sequence 2 (c) Sequence 3

Figure 5: Trajectories of the different sequences: the colouring from blue to red is used for identifying different time instants (blue - start, red - finish); the black dot represents the camera position. Top: top view; bottom: perspective view.



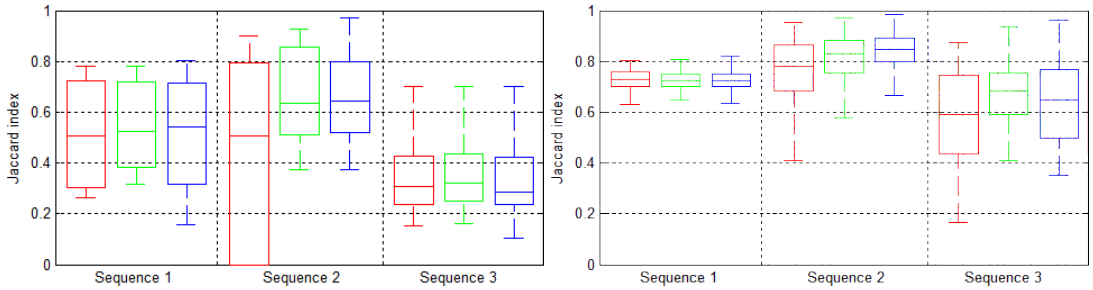
(a) Sequence 1 (b) Sequence 2 (c) Sequence 3

Figure 6: Tracking accuracy - Constant target size



(a) Sequence 1 (b) Sequence 2 (c) Sequence 3

Figure 7: Tracking accuracy - With adaptive scaling



(a) Constant target size scaling (b) With adaptive scaling

Figure 8: Tracking accuracy - Red: RGB, green: RGB-T, blue: RGB-D-T

REFERENCES

Bouguet, J.-Y. (2004). Camera calibration toolbox for matlab.

Choi, C. and Christensen, H. I. (2013). Rgb-d object tracking: A particle filter approach on gpu. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1084–1091. IEEE.

Endres, F., Hess, J., Engelhard, N., Sturm, J., Cremers, D., and Burgard, W. (2012). An evaluation of the rgb-d slam system. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Jafari, O. H., Mitzel, D., and Leibe, B. (2014). Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 5636–5643. IEEE.

Kumar, S., Marks, T. K., and Jones, M. (2014). Improving person tracking using an inexpensive thermal infrared sensor. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 217–224. IEEE.

Luber, M., Spinello, L., and Arras, K. O. (2011). People tracking in rgb-d data with on-line boosted target models. In *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*.

Matsumoto, K., Nakagawa, W., Saito, H., Sugimoto, M., Shibata, T., and Yachida, S. (2015). Ar visualization of thermal 3d model by hand-held cameras. In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications*, pages 480–487.

Mogelmose, A., Bahnsen, C., Moeslund, T. B., Clapés, A., and Escalera, S. (2013). Tri-modal person re-identification with rgb, depth and thermal features. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 301–307. IEEE.

Nakagawa, W., Matsumoto, K., de Sorbier, F., Sugimoto, M., Saito, H., Senda, S., Shibata, T., and Iketani, A. (2014). Visualization of temperature change using rgb-d camera and thermal camera. In *Computer Vision-ECCV 2014 Workshops*, pages 386–400. Springer.

Nummiaro, K., Koller-Meier, E., and Van Gool, L. (2002). Object tracking with an adaptive color-based particle filter. In *Pattern Recognition*, pages 353–360. Springer.

Nummiaro, K., Koller-Meier, E., and Van Gool, L. (2003). An adaptive color-based particle filter. *Image and vision computing*, 21(1):99–110.

Pérez, P., Hue, C., Vermaak, J., and Gangnet, M. (2002). Color-based probabilistic tracking. In *Computer visionECCV 2002*, pages 661–675. Springer.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*.

Stolkin, R., Rees, D., Talha, M., and Florescu, I. (2012). Bayesian fusion of thermal and visible spectra camera data for region based tracking with rapid background adaptation. In *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on*, pages 192–199. IEEE.

Susperregi, L., Martínez-Otzeta, J. M., Ansuategui, A., Ibarguren, A., and Sierra, B. (2013). Rgb-d, laser and thermal sensor fusion for people following in a mobile robot. *Int. J. Adv. Robot. Syst.*

Talha, M. and Stolkin, R. (2012). Adaptive fusion of infrared and visible spectra camera data for particle filter tracking of moving targets. In *Sensors, 2012 IEEE*, pages 1–4. IEEE.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*.

Vidas, S., Lakemond, R., Denman, S., Fookes, C., Sridharan, S., and Wark, T. (2012). A mask-based approach for the geometric calibration of thermal-infrared cameras. *Instrumentation and Measurement, IEEE Transactions on*, 61(6):1625–1635.

Vidas, S., Moghadam, P., and Bosse, M. (2013). 3d thermal mapping of building interiors using an rgb-d and thermal camera. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 2311–2318. IEEE.