Multi-Modal Tracking by Identification

Rok Mandeljc, Stanislav Kovačič, Matej Kristan, Janez Perš

Machine Vision Laboratory, Faculty of Electrical Engineering,
University of Ljubljana, Tržaška 25, 1000 Ljubljana, Slovenia
E-mail: rok.mandeljc@fe.uni-lj.si

Abstract

In this paper, we demonstrate, by performing quantitative evaluation, the benefit of tracking by identification over state-of-the-art identification by tracking. We evaluate four localization and tracking systems: a commercial localization system based on radio technology, a state-of-the-art computer-vision algorithm that uses multiple calibrated cameras to perform identification by tracking, and two multi-modal tracking-by-identification systems that have been developed in our laboratory. We briefly describe all four systems and evaluation metric, and present evaluation on a challenging indoor dataset.

1 Introduction

The need for recovery of individuals' *positions* and trajectories, measured in the *world coordinate system*, can be found in different applications, with most notable examples being closed-world surveillance, for example in high-security environments, where number of people and their visual appearance are a-priori known, analysis of athletes' performance in sports and sports medicine. Therefore, this problem is already a hot research topic and various localization solutions based on different sensor modalities have been proposed [1]. Two most prominent research areas are detection and tracking using video cameras [2] and localization using radio technology [3]. Sensor fusion has also become a prominent paradigm for overcoming limitations of individual sensor modalities [1].

Person detection and tracking using multiple cameras with overlapping fields of view has a long tradition in the field of computer vision. The existing approaches can be roughly divided into two groups. The so-called *detection-by-tracking* approaches are based on sequential techniques, such as Kalman filter or particle filters (e.g. [4]). While fast, and therefore considered state-of-the-art in the real-time tracking, recursive tracking may result in unbounded detection error propagation, which may eventually cause all targets to be lost. To mitigate this, the *tracking-by-detection* approaches employ robust frame-by-frame detection [5], on top of which optimization methods are applied for tracking [6], usually in offline and in batch manner. However, when it comes to maintaining the identities of tracks, these approaches perform *identification by tracking*; they rely on identity propagation along the track, with none or limited appearance validation. As such, they are prone to propagation of *identity switches* that occur when people come close. After people disperse, a propagated identity switch manifests as a large localization error, and essentially renders an individual's trajectory invalid.

In this paper, we demonstrate, by performing quantitative evaluation, the benefit of *tracking by identification* over state-of-the-art identification by tracking. We evaluate four localization and tracking systems: a commercial localization system based on radio technology (*System 1*), a state-of-the-art computer-vision algorithm that uses multiple calibrated cameras to perform identification by tracking (*System 2*), and two multi-modal tracking-by-identification systems (*System 3* and *System 4*), which have been developed in our laboratory. The systems are briefly described in Section 2, followed by description of evaluation metric in Section 3, experimental results with discussion in Section 4 and conclusion in Section 5.

2 Person detection and localization systems

2.1 System 1: radio-based tracking by identification

The radio-based localization system that we used is a commercially-available solution from Ubisense [7]. The system comprises a network of radio receivers and small radio-emitting tags that are worn by people. Localization is done based on time-difference-of-arrival (TDoA) and angle-of-arrival (AoA) [3] of radio pulses coming from tags, which in theory enables precise and robust localization. To prevent interference between tags, each tag is given time slots during which it emits signals; the effective frequency of the system is therefore 1–10 *Hz*, depending on the number of tags. For each tag, its position and unique identifier are returned by the system. Since identity is encoded in the radio signal, this system inherently does not suffer from identity switches and it can be considered as a tracking-by-identification system. However, as shown later on, there can be significant localization error caused by presence of radio-reflective metallic surfaces and obstacles in the signal's path.

2.2 System 2: camera-based identification by tracking

The second system used in our evaluation uses multiple calibrated video cameras with overlapping fields of view,

and consists of two parts. The first is Probabilistic Occupancy Map (POM) [5], the state-of-the-art person detection and localization algorithm. It involves discretizing the area of interest into a grid and estimating the probability of occupancy for each cell. It uses a simple and robust model that approximates silhouette at each location with a rectangle of fixed height for back-projecting currently-estimated probabilities back into views. The probability field (occupancy map) is then iteratively optimized so that the difference between input binary images, obtained from background subtraction, and back-projected images is minimized. The cells with sufficiently high probability of occupancy are considered to be occupied.

The anonymous detections obtained by POM on frame-by-frame basis are then linked into trajectories using K-Shortest Paths (KSP) tracker [6]. This is state-of-the-art *identification-by-tracking*, since linking is based solely on spatio-temporal proximity. The obtained trajectories inherently have no identities; they must be assigned manually, for example using the identity of a person that the trajectory was initialized on. Linking is performed off-line in batch manner, therefore the location data has same frequency as the input video stream.

2.3 Systems 3 and 4: multi-modal tracking by identification

Both *tracking-by-identification* systems use some of aforementioned components and are multi-modal: the first system uses both information from video cameras and the radio system, whereas the second system relies on multiple features that are obtained solely from video cameras.

In the first system, the detections from radio system are first anonymously integrated into the POM algorithm to improve anonymous detections [8]. The obtained anonymous detections are assigned identities from radio tags using spatially-optimal assignment obtained by Hungarian method [9]. Finally, each set of identified detections is separately linked into trajectories using KSP algorithm, which thereby performs tracking-by-identification.

The second system is based on fusion of multiple weak cues for detection (background subtraction, optical flow, detection of head and shoulders) and identification (color of shirt, height, prior knowledge about location) of individuals, which are encoded into a stack of feature maps [10, 11]. Multiple classifiers, one for each individual, are separately trained on an annotated part of video sequence, and then used to classify each cell as either empty or occupied by a specific individual. Due to limited amount of discriminative information, multiple classifiers can fire on a single cell. Again, we run KSP separately on top of the output of each classifier to link the identified detections.

3 Evaluation metric

At a given time instant (e.g. a video frame), the localization system returns D detection hypotheses, comprising their x and y coordinates on ground plane and unique identifier γ: $\mathbf{d}_i = (x_i^d, y_i^d, \gamma_i^d)$; $i = 1 \dots D$. Similarly, we have G ground-truth points: $\mathbf{g}_j = (x_j^g, y_j^g, \gamma_j^g)$; $j = 1 \dots G$. We construct cost matrix C, whose elements c_{ij} describe the cost of assigning each detection d_i to each ground truth point g_j, and find the optimal assignment using Hungarian method [9]. After assignment is done, the unassigned detections are considered to be *phantoms* (false positive detections) and unassigned ground truth points are considered *missing* (false negative) *detections*. From their count, we compute *precision*, *recall* and *F-score* and from *distances between assigned pairs* we obtain statistics about *2-D localization error*.

The cost function c_{ij} is primarily based on Euclidean distance between a detection and ground truth point, with some additional constraints. The value of ∞ prevents assignment of a detection and a ground truth point. We use three different cost functions, each resulting in an assignment that reveals a different aspect of the system:

$$c_{ij}^1 = \|\mathbf{d}_i - \mathbf{g}_j\| = \sqrt{(x_i^d - x_j^g)^2 + (y_i^d - y_j^g)^2} \quad (1)$$

$$c_{ij}^2 = \begin{cases} \|\mathbf{d}_i - \mathbf{g}_j\| & \text{if } \gamma_i^d = \gamma_j^g, \\ \infty & \text{otherwise} \end{cases} \quad (2)$$

$$c_{ij}^3 = \begin{cases} \|\mathbf{d}_i - \mathbf{g}_j\| & \text{if } \|\mathbf{d}_i - \mathbf{g}_j\| \leq T_d, \\ \infty & \text{otherwise} \end{cases} \quad (3)$$

When cost function c_{ij}^1 is used, each detection is assigned to its closest ground truth point, regardless of their identities. The resulting assignment therefore analyzes system's *detection and localization* performance. With c_{ij}^2, each detection is assigned to the closest ground truth point with same identity, which can, in case of identity switches, result in relatively large localization error; this can be used to analyze system's performance at *identification* or preserving the identities. A detailed analysis can be performed using c_{ij}^3. It assigns detections to their closest ground truth points, disregarding their identities. However, assignment is impossible unless distance is smaller than threshold T_d, which in our case is 0.5 *m*; therefore gross errors in detection and localization contribute to false positive and false negative detections and are not included in 2-D localization error statistics. Furthermore, once assignment is done, identities are compared and identity mismatches are summarized in a table, such as Table 1.

4 Results and discussion

In order to perform the evaluation, we captured a challenging dataset. We used four calibrated video cameras and Ubisense sensors to capture 3 minutes of data (video streams at 20 frames per second and events from the radio system) involving five people walking around a 7.1×7.0 *m* room (Figure 1). The setup is challenging both for radio-based and camera-based system, due to occlusions and presence of radio-reflective surfaces. Additionally, individuals are visually similar; three wear black and two wear grey clothes. For POM, we discretized room into 50×50 *cm* cells, placed 10 *cm* apart, and set the height of rectangles to 1.75*m*. Ground truth was annotated manually in each frame. Every 10th frame of the first minute

Figure 1: Views from each of the four cameras.

was used to train System 4, and all frames of the remaining two minutes were used for testing all four systems.

The 2-D localization error is analyzed by plotting its cumulative density function (CDF). Figures 2a, 2b and 2c show plots for assignments obtained for all four systems by using cost functions c_{ij}^1, c_{ij}^2 and c_{ij}^3, respectively. The detailed results obtained using cost function c_{ij}^3 are summarized in Table 1.

Comparing the curves on Figures 2a and 2b, we can see that performance of System 2, while good when only detection and localization are considered, drastically degrades when identity is considered as well. This is due to propagated identity switches, which also contribute to non-diagonal values in Table 1. On the other hand, the curves for identity-based systems remain unchanged; the systems exhibit same performance under both cost functions. The best-performing system is System 3, as it combines the best of its components: precise camera-based localization and reliable radio-based identities. It should be noted that KSP tracker, which is used in Systems 2, 3 and 4, tends to drift around during prolonged periods of co-occurring missing and phantom detections in the input data, resulting in gross localization errors. This is the main source of missing and phantom detections for Systems 3 and 4, as seen in Table 1, and is especially noticeable in case of Person #3, who relatively often fails to be detected by System 4. We therefore expect that integrating more visual features into System 4, and thereby improving the simultaneous detection and recognition, would make the system's performance comparable to that of System 3, but without people having to wear radio tags.

5 Conclusion

We presented a quantitative evaluation of four person detection and localization systems: a commercially available radio-based system, a system based on state-of-the-art computer-vision algorithm that performs identification by tracking, and two multi-modal systems that perform tracking by identification. Our experiment shows that tracking-by-identification gives better results, as it can prevent propagation of identity switches. The best-performing system combines precise camera-based localization and reliable identity information coming from the radio-based system. However, this comes at the price of people having to wear radio tags, which is possible only in certain scenarios. Therefore, our future work will focus on improving the multi-modal system that relies only on visual cues and whose unobtrusive nature should allow deployment in wider range of applications.

Acknowledgements

This work was supported by the research program P2-0095, research project J2-4284 and the research grant 1000-10-310118, all by Slovenian Research Agency.

References

[1] J. Hightower and G. Borriello, "Location systems for ubiquitous computing," *Computer*, vol. 34, pp. 57–66, aug. 2001. 1

[2] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, December 2006. 1

[3] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE TSMC, Part C*, vol. 37, pp. 1067–1080, nov. 2007. 1

[4] M. Kristan, J. Perš, M. Perše, and S. Kovačič, "Closed-world tracking of multiple interacting targets for indoor-sports applications," *CVIU*, vol. 113, no. 5, pp. 598 – 611, 2009. 1

[5] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE TPAMI*, vol. 30, no. 2, pp. 267–282, 2007. 1, 2

[6] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using K-Shortest paths optimization," *IEEE TPAMI*, vol. 33, no. 9, pp. 1806–1819, 2011. 1, 2

[7] "Research & Development Packages — Ubisense." `http://www.ubisense.net/en/rtls-solutions/research-packages.html`. Accessed: 07/2012. 1

[8] R. Mandeljc, J. Perš, M. Kristan, and S. Kovačič, "Fusion of non-visual modalities into the probabilistic occupancy map framework for person localization," in *Proc. of IEEE Int. Conf. on Distributed Smart Cameras (ICDSC2011)*, pp. 1–6, aug. 2011. 2

[9] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics (NRL)*, vol. 52, no. 1, pp. 7–21, 2005. 2

[10] R. Mandeljc, J. Perš, M. Kristan, and S. Kovačič, "Detekcija, lokalizacija in identifikacija oseb z več kamerami ter mapami značilnic," in *ROSUS2012*, pp. 1–7, mar. 2012. 2

[11] R. Mandeljc, S. Kovačič, M. Kristan, and J. Perš, "Non-sequential multi-view detection, localization and identification of people using multi-modal feature maps." Submitted to *ACCV2012*. 2

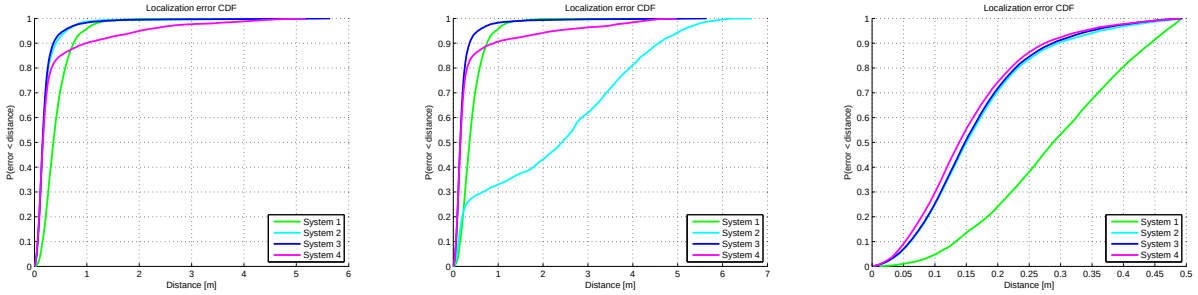
Table 1: Analysis of results for all four systems, using cost function c_{ij}^3 (Eq. 3). Each detection column sums up to number of detections with that particular identity and each object row sums up to number of ground truth points for that object; dividing diagonal elements by column sums yields precision, while dividing by row sums yields recall.

SYSTEM 1		**DETECTIONS**						**Precision**	**Recall**	**F-Score**
		#1	**#2**	**#3**	**#4**	**#5**	**Missing**			
OBJECT	**#1**	**1968**	5	38	0	35	575	0.7520	0.7509	0.7514
	#2	1	**2020**	0	0	0	600	0.7707	0.7707	0.7707
	#3	7	9	**1306**	0	12	1287	0.5000	0.4983	0.4991
	#4	0	0	0	**1937**	8	676	0.7390	0.7390	0.7390
	#5	30	2	39	2	**1962**	586	0.7497	0.7486	0.7491
Phantom		611	585	1229	682	600	—	**0.7024**	**0.7015**	**0.7019**

SYSTEM 2		**DETECTIONS**						**Precision**	**Recall**	**F-Score**
		#1	**#2**	**#3**	**#4**	**#5**	**Missing**			
OBJECT	**#1**	**1149**	652	278	115	149	278	0.4384	0.4384	0.4384
	#2	0	**553**	749	448	824	47	0.2110	0.2110	0.2110
	#3	4	895	**39**	1428	148	107	0.0149	0.0149	0.0149
	#4	1181	0	635	**484**	0	321	0.1847	0.1847	0.1847
	#5	0	278	699	29	**1430**	185	0.5456	0.5456	0.5456
Phantom		287	243	221	117	70	—	**0.2789**	**0.2789**	**0.2789**

SYSTEM 3		**DETECTIONS**						**Precision**	**Recall**	**F-Score**
		#1	**#2**	**#3**	**#4**	**#5**	**Missing**			
OBJECT	**#1**	**2356**	2	26	0	20	217	0.8989	0.8989	0.8989
	#2	7	**2535**	5	0	15	59	0.9672	0.9672	0.9672
	#3	36	18	**2487**	0	4	76	0.9489	0.9489	0.9489
	#4	0	0	0	**2421**	3	197	0.9237	0.9237	0.9237
	#5	16	17	0	1	**2402**	185	0.9164	0.9164	0.9164
Phantom		206	49	103	199	177	—	**0.9310**	**0.9310**	**0.9310**

SYSTEM 4		**DETECTIONS**						**Precision**	**Recall**	**F-Score**
		#1	**#2**	**#3**	**#4**	**#5**	**Missing**			
OBJECT	**#1**	**2472**	35	5	11	7	91	0.9432	0.9432	0.9432
	#2	0	**1976**	8	26	14	597	0.7539	0.7539	0.7539
	#3	15	39	**1742**	0	2	823	0.6646	0.6646	0.6646
	#4	3	86	0	**2359**	0	173	0.9000	0.9000	0.9000
	#5	1	42	207	0	**2280**	91	0.8699	0.8699	0.8699
Phantom		130	443	659	225	318	—	**0.8263**	**0.8263**	**0.8263**



(a) cost function c_{ij}^1 (Eq. 1) (b) cost function cost c_{ij}^2 (Eq. 2) (c) cost function c_{ij}^3 (Eq. 3)

Figure 2: Cumulative density function of 2-D localization error for assignments constructed using different costs and all four systems: System 1 (radio-based Ubisense), System 2 (KSP tracker on top of anonymous POM detections), System 3 (KSP tracker on top of POM detections with identities from radio system) and System 4 (KSP tracker on top of fusion of multiple visual cues for detection and identification).