Accurate Fusion of Robot, Camera and Wireless Sensors for Surveillance Applications

Andrew Gilbert, John Illingworth and Richard Bowden
CVSSP, University of Surrey, Guildford, Surrey GU2 7XH United Kingdom
a.gilbert@surrey.ac.uk

Jesus Capitan, University of Seville, Seville, Spain
Luis Merino, Pablo de Olavide University, Seville, Spain
merino@cartuja.us.es

Abstract

Often within the field of tracking people only fixed cameras are used. This can mean that when the the illumination of the image changes or object occlusion occurs, the tracking can fail. We propose an approach that uses three simultaneous separate sensors. The fixed surveillance cameras track objects of interest cross camera through incrementally learning relationships between regions on the image. Cameras and laser rangefinder sensors onboard robots also provide an estimate of the person. Moreover, the signal strength of mobile devices carried by the person can be used to estimate his position. The estimate from all these sources are then combined using data fusion to provide an increase in performance. We present results of the fixed camera based tracking operating in real time on a large outdoor environment of over 20 non-overlapping cameras. Moreover, the tracking algorithms for robots and wireless nodes are described. A decentralized data fusion algorithm for combining all these information is presented.

1. Introduction

Surveillance cameras are increasingly being used as a tool to monitor and deter crime. As a result, there are large numbers of cameras which lack effective continuous monitoring due to the limitations of humans in managing large-scale systems. Therefore, tools to assist and aid the operator's decision process are essential. The approach presented in this paper aims to automatically track objects in (intra) and between (inter) cameras. This can be a challenging problem using a single sensor such as fixed cameras. Therefore a number of individually weaker sensors are used, these are then combined with an advanced data fusion technique.

The fixed cameras are situated outside within the *Universitat Politécnica de Catalunya* in Barcelona. The cameras have been installed for the EU Project called URUS ((**U**biquitous Networking **R**obotics in **U**rban **S**ettings) [**?**], the project aims to combine advances within the vision and robotics fields to allows for complex guidance and evacuation scenarios to be implemented. At the beginning of the scenario a person is to be detected as waving for attention and then must be tracked continually to allow a robot as in Figure 1 to go and guide him. The cameras are assumed to be non overlapping and are all outside. This means that an image based tracker would individually be insufficient to track objects of interest.

Figure 1. An example of Guiding, interaction and transportation of people and goods.

Therefore, the system will consider, besides the data from from the cameras, information from a number of other sensors, namely cameras and laser rangefinders on board the robot and a Wireless Sensor Network (WSN). All these elements can communicate through wireless links, and constitute what it is called a Network Robot System (NRS). A NRS is a new concept that integrates robots, sensors, communications and mobile devices in a cooperative way, which means not only a physical interconnection between these elements, but also, for example, the development of novel intelligent methods of cooperation for task oriented purposes [?]. For instance, fusing all the information from all the elements in the NRS will allow for a more accurate tracking of a person in missions such as human guidance by a robot in a urban environment.

This paper will first present the individual input sensor algorithms. The fixed camera tracking algorithm learns the relationships between the non-overlapping cameras automatically. This is achieved by modelling the colour, and temporal relationship between cameras. The two cues are deliberately very weak as more detailed and complex cues would not be able to work with the low resolution and real time requirements of the system. These two cues are then used to weight a simple appearance likelihood of the person for inter camera racking.

Robots and the WSN also provide observations on the person being tracked. The paper summarizes the main ideas on the tracking from robots, and on the use of the signal strength from wireless sensors for tracking. Finally, the results of the tracking from all sensors are used to infer the position of the person in a global coordinate system through a data fusion process. In order to cope with scalability, a decentralized data fusion system is employed to combine the information provided by all the systems to have a consistent and improved estimation of the person's position at any given moment. The decentralized data fusion process can obtain the same results than a ideal centralized systems that receives all the data at any given moment, but being more robust under delays and communication failures and scalable with the size of the system.

2. Related Work

There has been many attempts to track people and other moving objects inter camera. The early tracking algorithms [?, ?, ?, ?] required both camera calibration and overlapping fields of view. Others [?, ?] can work with non-overlapping cameras but still require calibration. Probabilistic approaches have been presented [?, ?], however these are often limited in application due to restrictive assumptions. KaewTraKulPong and Bowden [?] or Ellis *et al* [?] do not require *a priori* correspondences to be explicitly stated, instead they use the observed motion over time to establish reappearance periods. In both cases batch processing was performed on the data which limits their application.

In this paper, a system that employs not only the fixed cameras, but information from robots and wireless sensors is used as well. There are an increasing interest in tracking systems that use signal strength received by mobile devices [?, ?]. However, most systems are devoted to the localization of static devices. Moreover, there are systems for tracking persons using robots and vision [?]. One contribution of this paper is the combination of these sources for cooperative tracking.

There are many systems that employ a central server that receives all the information to obtain a fused estimation of the quantity to be estimated, like the person position [?]. However, these kind of systems are dependant on this central node, are not robust under communications failures, latencies or drop outs, and do not scale well with the number of nodes. In this case, it is preferable to have a decentralized system in which each part only employs local information and exchange with its peers its local estimation. The main issues and problems with decentralized information fusion can be traced back to the work [?], where the Information Filter (IF, dual of the Kalman Filter) is used as the main tool for data fusion for process plant monitoring. The IF has very nice characteristics for decentralization, and for instance it has been used for decentralized perception with aerial robots in [?, ?]. However, in these cases, the filters are suboptimal, as they cannot recover the same estimation than a central node in the case of tracking scenarios.

3. URUS System overview

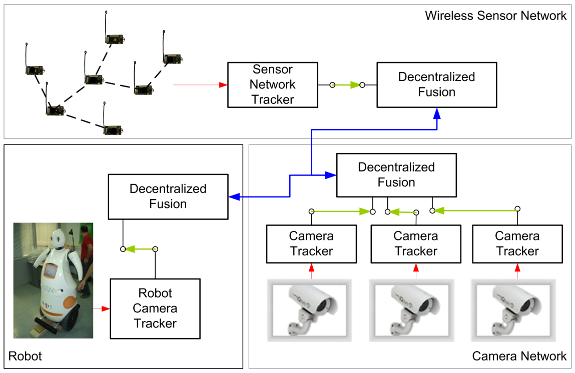


Figure 2. A block description of the URUS system

The NRS developed in the URUS Project consists of a team of mobile robots, equipped with cameras, laser rangefinders and other sensors for localization, navigation and perception; a fixed camera network of more than 20 cameras for environment perception; and a wireless sensor network, that uses the signal strength of the received messages from a mobile device to determine the position of a person carrying it. Figure 2 shows an scheme of the system and Figure 3 two frames from one of the cameras.

From the perception point of view, in the NRS the information obtained by the fixed camera network or the wireless sensor network can be shared with the information each robot obtains about the environment to improve the perception, so that each robot obtains a better picture of the world

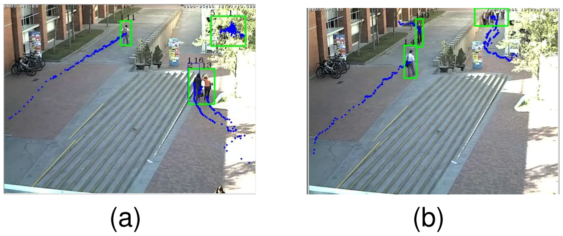


Figure 3. (a)Frame 50 on camera 5, (b) Frame 450 on camera 5

than if it would be alone, for instance for the task of person guiding. In this case, the tracks on the image plane obtained by the camera network will be combined with the information of the other systems (robots and WSN) to improve the tracking of the person being guided. This way, it is possible to cope with occlusions, obtain better tracking capabilities as information of different modalities is employed, and non covered zones, as the robots can move to cover these zones.

4. Fixed Camera Tracking

The fixed cameras cover a wide area of the experiment site and are the basis for the fusion of the other sensors, they are able to track objects of interest both on and across different cameras without explicit calibration.

The approach is based on the method proposed by Gilbert and Bowden previously [?]. Intra camera objects of interest are identified with a Gaussian mixture model [?] and are linked temporally with a Kalman filter to provide movement trajectories intra camera. When the object of interest enters a new camera, the transfer of the object to the new camera is a challenge as cameras have no overlapping fields of view, making many traditional image plane calibration techniques impossible. In addition the large number of cameras mean traditional time consuming calibration is infeasible. Therefore the approach needs to learn the relationships between the cameras automatically. This is achieved by the way of two cues, modelling the colour, and movement of objects inter camera. These two weak cues, are then combined to allow the technique to determine if objects have been previously tracked on another camera or are new object instances. The approach learns these camera relationships, though unlike previous work does not require *a priori* calibration or explicit training periods. Incrementally learning the cues over time allows for the accuracy to increase without any supervised input.

4.1. Forming Temporal links between Cameras

To learn the temporal links between cameras, we make use of the key assumption that, given time, objects (such as people) will follow similar routes inter camera and that the repetition of the routes will form marked and consistent trends in the overall data.

Initially the system is subdivided so that each camera is a single region. It identifies temporal reappearance links at the camera-to-camera level. After sufficient evidence has been accumulated, the noise floor level is measured for each link. If the maximum peak of the distribution is found to exceed the noise floor level, this indicates a possible correlation between the two blocks as shown in Figure 4). If a link is found between two regions, they are both subdivided to each create four new equal sized regions providing a higher level of detail. While regions with little or no data are removed to maintain scalability.

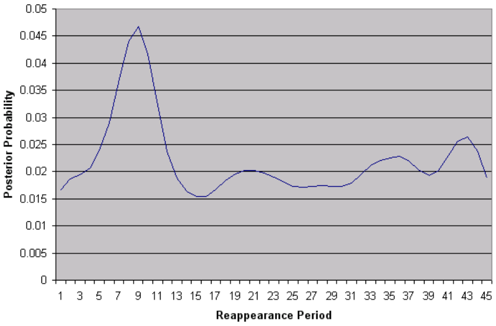


Figure 4. An example of a probability distribution showing a distinct link between two regions

4.2. Modelling Colour Variations

The colour quantisation descriptor used to form temporal reappearance links in the previous section, assumes a similar colour response between cameras. However this is seldom the case. Therefore, a colour calibration of these cameras is proposed that can be learnt incrementally simultaneously with the temporal relationships discussed in the section above. People tracked inter camera are automatically used as the calibration objects, and a transformation matrix is formed incrementally to model the colour changes between specific cameras.

The transformation matrices for the cameras are initialised as identity matrices assuming a uniform prior of colour variation between camera. When a person is tracked inter camera and is identified as the same object, the difference between the two colour descriptors, is modelled by a transform matrix . The matrix is calculated by computing the transformation that maps the person's descriptor from the previous camera to the person's current descriptor. This transformation is computed via SVD. The matrix is then averaged with the appropriate camera transformation matrix, and repeated with other tracked people to gradually build a colour transformation between cameras.

4.3. Calculating Posterior Appearance Distributions

With the weak cues learnt, when an object which leaves a in region y we can model its reappearance probability over time as;

$$P(O_t|O_y) = \sum_{\forall x} w_x P(O_{x,t}|O_y) \tag{1}$$

where the weight w_x at time t is given as

$$w_x = \frac{\sum_{i=0}^{T} f_{\phi}^{x|y}}{\sum_{\forall y} \sum_{i=0}^{T} f_{\phi}^{x|y}} \tag{2}$$

This probability is then used to weight the observation likelihood obtained through colour similarity to obtain a posterior probability of a match. Tracking objects is then achieved by maximising the posterior probability within a set time window.

5. Robot Cameras Tracking

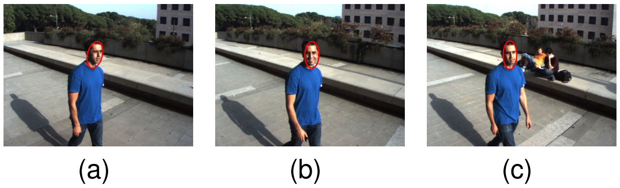


Figure 5. People tracking from robot camera for guiding

The robots used carry on board cameras that are used for person guiding. The algorithms employed for this are based on a combination of person detection and tracking. The tracking algorithm is based on the mean shift technique [**?**]. In parallel, a face detection algorithm is applied to the image [**?**], the results from the tracking and the detection applications are combined, so that the robot employs the face detector when the tracker is lost to recover the track. Some improvements can be applied to the features in order to cope with illuminations changes [**?**]. As a result, the robots can obtain estimations of the pose of the person face on the image plane (see Fig. 5).

6. Wireless Sensor Network

In the NRS, a network of wireless Mica2 sensor nodes are used. These Mica2 nodes are able to sense different quantities, like pressure, temperature, humidity, etc. Moreover, they have wireless communication devices, and are able to form networks and relay the information they gather to a gateway. In addition the signal strength received by the set of static nodes (Received Signal Strength Indicator, RSSI) can be used to infer the position of a mobile object or a person carrying one of the nodes.

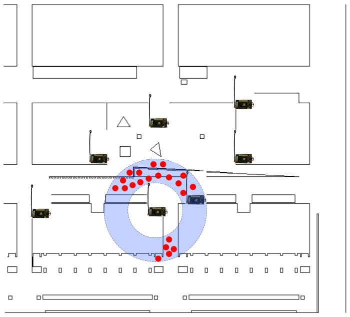


Figure 6. The signal received by a set of static nodes can be used to infer the position of the node. Particles (red) are used to represent person hypotheses.

The algorithm to estimate the node position is based on particle filtering. In the particle filter, the current belief about the position of the mobile node is represented by a set of particles that represent hypotheses about the actual position of the person that carries the node (see Fig. 6).

In each iteration of the filter, kinematic models of the motion of the person and map information are used to predict the future position of the particles. The addition of the map information identify and discard impossible motions.

When new messages are received by the network of static nodes, the weight of the different particles is updated by considering the RSSI indicator. By considering radio propagation models, it is possible to determine the likelihood of receiving a certain signal power by considering the distance from the particle to the receiving node [**?**]. Each transmission restricts the position of the particles to a annular shaped area around the receiving node (see Fig. 6).

As a result, the filter can provide estimations on the 3D position of the mobile node with a one metre accuracy. Figure 7 shows the evolution of the particles for a particular guiding experiment at the Barcelona fixed camera experiment site. Figure 8 shows the estimated position of the person estimated by the WSN, compared to that of the guiding robot (that is some metres ahead).

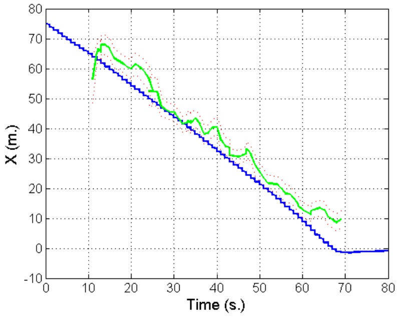


Figure 8. Estimated position of the person by the WSN (green) and position of the guiding robot (blue) estimated by using GPS.

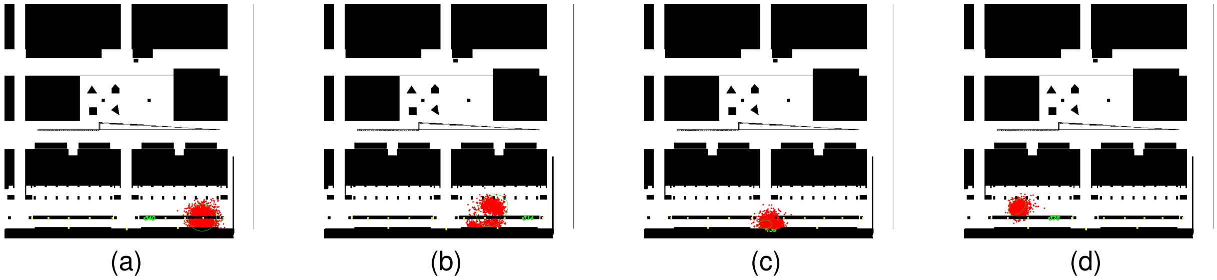


Figure 7. A sequence of the particles employed in the filter for this experiment

7. Decentralized Data Fusion for Person Tracking

Using the trackers described above, the camera network and the robots will be able to obtain local estimations of the position of the persons on the image plane. These estimations, characterized as Gaussian distributions (mean and covariance matrix) and the ones provided by the WSN, can be fused in order to obtain an accurate estimation of the 3D position of the person.

One potential solution is to have a central node that implements a centralized Extended Kalman Filter (EKF). However this filter have must access to all the estimations $\mathbf{z}^t = [\mathbf{z}_1^{tT}, \ldots, \mathbf{z}_M^{tT}]^T$ at any moment. However, a centralized system presents some drawbacks already commented.

The idea is to implement a decentralized estimation system, in that each node only employs local information (data only from local sensors, for instance, a camera subnet, or the sensors on board the robot), and then *shares* this information with other nodes (see Fig. 2). The Information Filter (IF), which corresponds to the dual implementation of the Kalman Filter (KF), is very suitable for a decentralized estimation. While the KF represents the distribution using its first $\boldsymbol{\mu}$ and second $\boldsymbol{\Sigma}$ order moments, the IF employs the so-called *canonical representation*. The fundamental elements are the *information vector* $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and the *information matrix* $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Prediction and updating equations for the (standard) IF can also be derived from the standard KF [**?**]. In the case of non-linear prediction or measurement, first order linearisation leads to the Extended Information Filter (EIF).

Let us consider the system:

$$\mathbf{x}_t = \mathbf{A}_t \mathbf{x}_{t-1} + \boldsymbol{\nu}_t \qquad (3)$$

$$\mathbf{z}_t = \mathbf{g}_t(\mathbf{x}_t) + \boldsymbol{\varepsilon}_t \qquad (4)$$

where $\mathbf{x}_t$ is the person's position and velocity at time t and $\mathbf{z}_t$ represents the estimations obtained by the camera network, robots and WSN at time t, and $\mathbf{g}_t$ is the measurement function (for instance, the pin-hole model of the cameras). Knowing the information matrix and vector for the person trajectory up to time $t-1$, $\boldsymbol{\Omega}^{t-1}$ and $\boldsymbol{\xi}^{t-1}$, the EIF for update the estimated trajectory of the person in a centralized system is given by Algorithm 1, where $\mathbf{M}_t = \nabla \mathbf{g}_t(\bar{\boldsymbol{\mu}}_t)$, $\mathbf{R}_t$ is the covariance of the additive noise for the prediction model (3) and $\mathbf{S}_t$ is the covariance matrix of the estimations. **Add_M** adds a block row and a block column of zeros to the previous information matrix and **Add_Row** adds a block row of zeros to the previous information vector. The filter is initialized with an initial estimation of the person position.

```
Algorithm 1 $(\boldsymbol{\xi}^t, \boldsymbol{\Omega}^t) \leftarrow$Information Filter$(\boldsymbol{\xi}^{t-1}, \boldsymbol{\Omega}^{t-1}, \mathbf{z}_t)$
1: $\bar{\boldsymbol{\Omega}}^t = \mathbf{Add\_M}(\boldsymbol{\Omega}^{t-1}) + \begin{pmatrix} \begin{pmatrix} \mathbf{I} \\ -\mathbf{A}_t^T \end{pmatrix} \mathbf{R}_t^{-1} \begin{pmatrix} \mathbf{I} & -\mathbf{A}_t \end{pmatrix} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$
2: $\bar{\boldsymbol{\xi}}^t = \mathbf{Add\_Row}(\boldsymbol{\xi}^{t-1})$
3: $\boldsymbol{\Omega}^t = \bar{\boldsymbol{\Omega}}^t + \sum_j \begin{pmatrix} \mathbf{M}_{j,t}^T \mathbf{S}_{j,t}^{-1} \mathbf{M}_{j,t} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$
4: $\boldsymbol{\xi}^t = \bar{\boldsymbol{\xi}}_t + \sum_j \begin{pmatrix} \mathbf{M}_{j,t}^T \mathbf{S}_{j,t}^{-1} (\mathbf{z}_{j,t} - \mathbf{g}_t(\bar{\boldsymbol{\mu}}_t) + \mathbf{M}_t \bar{\boldsymbol{\mu}}_t) \\ \mathbf{0} \end{pmatrix}$
```

7.1. Decentralized Information Filter

The main interest of the IF is that it can be easily decentralized. In a decentralized approach, each node i of the network employs only its local data $\mathbf{z}_i^t$ to obtain a local estimation of the person trajectory (given by $\boldsymbol{\xi}^{i,t}$ and $\boldsymbol{\Omega}^{i,t}$) and then *shares* its belief with its neighbours. The received information from other nodes is locally fused in order to improve the local perception of the world. The decentralized fusion rule should, produce the same result locally that obtained by a central node employing a centralized EIF.

Therefore, each node will run Algorithm 1 using only its local information. When a node i is within communication range with other node j, they can share their beliefs, represented by their information vectors $\boldsymbol{\xi}^{i,t}$ and $\boldsymbol{\xi}^{j,t}$, and matrices $\boldsymbol{\Omega}^{i,t}$ and $\boldsymbol{\Omega}^{j,t}$. It can be seen [**?**] that the fusion rule is:

$$\mathbf{\Omega}^{i,t} \leftarrow \mathbf{\Omega}^{i,t} + \mathbf{\Omega}^{j,t} - \mathbf{\Omega}^{ij,t} \tag{5}$$

$$\boldsymbol{\xi}^{i,t} \leftarrow \boldsymbol{\xi}^{i,t} + \boldsymbol{\xi}^{j,t} - \boldsymbol{\xi}^{ij,t} \tag{6}$$

The equations mean that each node must sum up the information received from other nodes. This can be derived from the updating steps 3 and 4 of Algorithm 1, as each node computes part of the updating that a central node would compute (which is a sum for all the information received). The additional term $\mathbf{\Omega}^{ij,t}$ and $\boldsymbol{\xi}^{ij,t}$ represents the common information between the nodes. This common information is due to previous communications between nodes, and should be removed to avoid double counting of information (known as rumour propagation). This common information can be maintained by a a separated EIF called channel filter [**?**] to maintain $\mathbf{\Omega}^{ij,t}$ and $\boldsymbol{\xi}^{ij,t}$. This common information can be locally estimated assuming a tree-shaped network topology (no cycles or duplicated paths of information).

It is important to remark that, using this fusion equation and considering trajectories (delayed states), the local estimator can obtain an estimation that is equal to that obtained by a centralized system [**?**]. Another advantage of using delayed states is that the belief states can be received asynchronously. Each node in the NRS can accumulate evidence, and send it whenever it is possible. However, as the state grows over time, the size of the message needed to communicate its belief also does. For the normal operation of the system, only the state trajectory over a time interval is needed, so these belief trajectories can be bounded. Note that the trajectories should be longer than the maximum expected delay in the network in order not to miss any measurements information.

8. Experimental Results

A series of experiments were performed on the fixed camera system described in section 3.

To illustrate the incrementally learnt cross camera calibration, the inter camera relationships were learnt for a total of 5 days. The number of tracked objects of interest on each camera was 200 per day. This is relatively low and made the region subdivision unsuitable after the second level of subdivision. Figure 9 shows resultant temporal likelihoods for a number of inter camera links at a single subdivision level.

The black vertical line indicates a reappearance of zero seconds, it can be seen that there is strong links between cameras 3 and 4 and between 3 and 5. While there is no visible link between 3 and 6 and between 3 and 14. This is due to the increased distance and people will rarely reappear on cameras 6 and 14 after they were tracked on camera 3.

Table 1 shows the accuracy results of tracking people inter camera. The inter camera links were formed over up to 5 days and the test sequence consists of a 1 hour sequence on the cameras, with a total of 50 people tracked inter camera. A 1 subdivision is a region per camera, 2 subdivision is the where any linked regions are subdivided as described in section 4.1. All people that moved inter camera were groundtruthed and a true positive occurred when a person was assigned the same ID as that they were assigned on a previous camera.

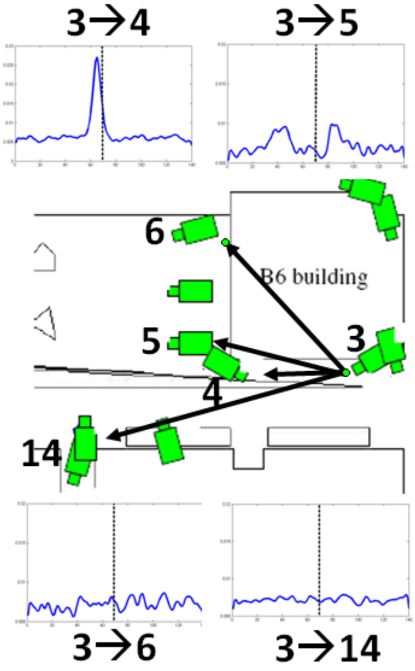
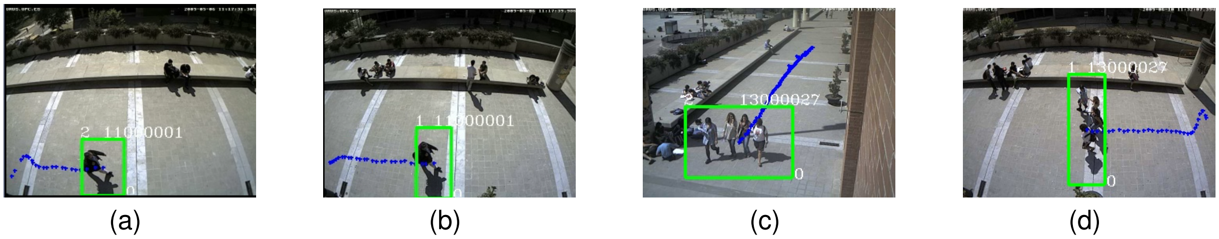


Figure 9. Inter camera temporal likelihoods

Table 1.

	Data Amount (days)			
Method	0	1	2	5
1 Subdiv	34%	38%	56%	78%
2 Subdiv	34%	10%	60%	83%

The column for 0 days indicates performance without learning the camera time and colour relationships. It is poor generally due to large colour variations inter camera due to shadow and lighting changes. The 2 level subdivision initially performs poorly as it requires greater data to build relationships. However by 5 days significant improvement is shown for both one and two levels of region subdivision. Little performance is gained from the additional subdivision on this system due to the lower levels of traffic and low level of routes between the cameras due to their closeness. However for a more distributed system the additional detail of the region relationships would aid the tracking performance greater. Figure 10 gives example frames of tracking inter camera for two separate people.



(a) (b) (c) (d)

Figure 10. Cross camera Tracking(a) Person 11000001 on camera 11, (b) Person 11000001 correctly identified on camera 12 (c) Person 13000027 on camera 13 (d) Person 13000027 correctly identified on camera 12.

8.1. Data fusion

In order to illustrate the benefits from the data fusion process, a simple setup is presented here. This setup consists of two fixed cameras and a WSN. The objective was to track one person cooperatively. In the experiment, the person is not always in the field of view of the cameras, appearing and disappearing from the image plane several times.

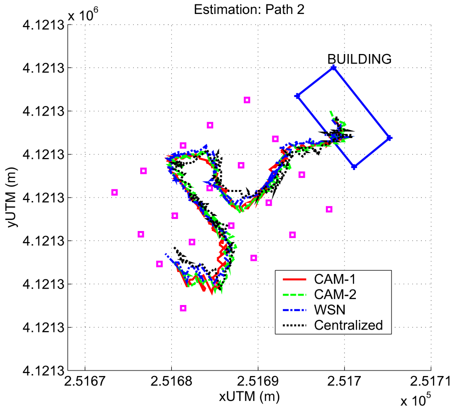


Figure 11. Estimated trajectory by the three elements of the NRS, cameras and WSN, compared to a centralized estimation. The elements converge to a very similar estimation, even if for some time the cameras are not seeing the person.

Three instances of the decentralized algorithm are launched, processing each camera's data and the WSN estimations. They communicate to exchange estimated trajectories, as described in section 7.1. The received data is fused into the local estimations, leading to a decentralized tracking of the person. Figure 11 shows the estimated trajectory of the person for the different elements, and the trajectory estimated by a centralized node with access to all information for comparison.

Figure 12 shows the X and Y estimations compared with the centralized estimation. One important benefit from the system is that, if the communications channels are active, the different elements have nearly the same information. That way, one robot or camera, even not seeing the person, can know where it is. Also, the uncertainty is decreased due to the redundancies in the system.

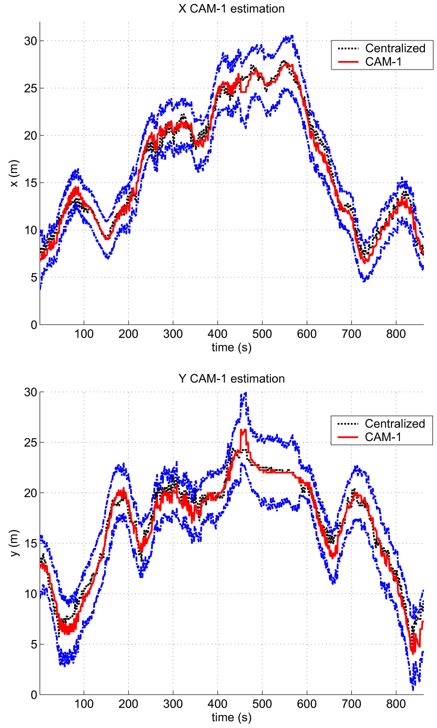


Figure 12. Estimated X and Y by the camera (red) compared to the centralized estimation (black), with the 3-sigma bound. The estimations are nearly the same, except during times when no communications occur.

9. Conclusion

The combination of robots and ambient intelligence (like embedded sensors and camera networks) seems a clear trend in the near future. This paper has presented a data fusion method that aims to use multiple sensors to accurately track people within a surveillance context. The algorithms are real time and operate on realistic outdoor environments. The fixed camera tracking provides high accuracy by incrementally learning the colour and temporal relationships between regions on non overlapping cameras. Moreover, the signal strength of mobile devices is employed to estimate the position of the person by using particle filtering. The combination of all this information with that obtained by robots allows accurate person tracking in more challenging situations. In the short future, the authors will test the system in more complex scenarios, involving several robots, 20 fixed cameras and a net of 30 wireless sensors in the missions of person guiding and surveillance.

10. Acknowledgements

This work is supported by URUS, Ubiquitous networking Robotics in Urban Settings, funded by the European Commission under FP6 with contract number FP6-EU-IST-045062, and CONET, the Cooperating Objects Network of Excellence, funded by the European Commission under FP7 with contract number FP7-2007-2-224053"

References

[1] F. Bourgault and H. Durrant-Whyte. Communication in general decentralized filters and the coordinated search strategy. In *Proc. of The 7th Int. Conf. on Information Fusion*, pages 723–730, 2004.

[2] G. R. Bradski. Computer vision face tracking for use in a perceptual user interface, 1998.

[3] F. Caballero, L. Merino, P. Gil, I. Maza, and A. Ollero. A probabilistic framework for entire wsn localization using a mobile robot. *Journal of Robotics and Autonomous Systems*, 56(10):798–806, 2008.

[4] Q. Cai and J. Aggarwal. "Automatic Tracking of Human Motion in Indoor Scenes across Multiple Synchronized Sideo Streams". *In Proc. of IEEE International Conference on Computer Vision (ICCV'98)*, 1998.

[5] J. Capitán, D. Mantecón, P. Soriano, and A. Ollero. Autonomous perception techniques for urban and industrial fire scenarios. In *Proceedings of IEEE International Workshop on Safety, Security, and Rescue Robotics (SSRR)*, pages 1–6, Rome, Italy, September 2007.

[6] J. Capitán, L. Merino, F. Caballero, and A. Ollero. Delayed-State Information Filter for Cooperative Decentralized Tracking. In *Proceedings of the International Conference on Robotics and Automation, ICRA*, 2009.

[7] T. Chang, S. Gong, and E. Ong. "Tracking Multiple People under Occlusion using Multiple Cameras". *In Proc. of BMVA British Machine Vision Conference (BMVC'00)*, pages 566–575, 2000.

[8] S. Dockstader and A. Tekalp. "Multiple Camera Tracking of Interacting and Occluded Human Motion". *In Proc. of IEEE*, 89(10):1441–1455, 2001.

[9] T. Ellis, D. Makris, and J. Black. "Learning a Multi-Camera Topology". *In Proc. of Joint IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 165–171, 2003.

[10] A. Gilbert and R. Bowden. "Incremental Modelling of the Posterior Distribution of Objects for Inter and Intra Camera Tracking ". *Proc. BMVC'05, Oxford UK*, 2005.

[11] S. Grime and H. F. Durrant-Whyte. Data fusion in decentralized sensor networks. *Control Engineering Practice*, 2(5):849–863, Oct. 1994.

[12] T. Huang and S. Russell. "Object Identification in a Bayesian Context". *In Proc. of International Joint Conference on Artificial Intelligence (IJCAI-97)*, pages 1276–1283, 1997.

[13] P. KaewTrakulPong and R. Bowden. "A Real-time Adaptive Visual Surveillance System for Tracking Low Resolution Colour Targets in Dynamically Changing Scenes". *In Journal of Image and Vision Computing*, 21(10):913–929, 2003.

[14] P. Kelly, A. Katkere, D. Kuramura, S. Moezzi, and S. Chatterjee. "An Architecture for Multiple Perspective Interactive Video". *In Proc. of the 3rd ACE International Conference on Multimedia*, pages 201–212, 1995.

[15] V. Kettnaker and R. Zabih. "Bayesian Multi-Camera Surveillance". *In Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'99)*, pages 253–259, 1999.

[16] V. Morariu and O. Camps. "Modeling Correspondences for Multi-Camera Tracking using Nonlinear Manifold Learning and Target Dynamics". *In Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'06)*, I:545–552, 2006.

[17] E. Nettleton, H. Durrant-Whyte, and S. Sukkarieh. A robust architecture for decentralised data fusion. In *Proc. of the International Conference on Advanced Robotics (ICAR)*, 2003.

[18] V. Ramadurai and M. L. Sichitiu. Localization in wireless sensor networks: A probabilistic approach. In *Proceedings of the 2003 International Conference on Wireless Networks (ICWN 2003)*, pages 275–281, Las Vegas, NV, June 2003.

[19] A. Sanfeliu and J. Andrade-Cetto. Ubiquitous networking robotics in urban settings. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2006*, 2006.

[20] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. People Tracking with Mobile Robots Using Sample-Based Joint Probabilistic Data Association Filters. *The International Journal of Robotics Research*, 22(2):99–116, 2003.

[21] S. Sukkarieh, E. Nettleton, J.-H. Kim, M. Ridley, A. Goktogan, and H. Durrant-Whyte. The ANSER Project: Data Fusion Across Multiple Uninhabited Air Vehicles. *The International Journal of Robotics Research*, 22(7-8):505–539, 2003.

[22] M. Trivedi, I. Mikic, and S. Bhonsle. "Active Camera Networks and Semantic Event Databases for Intelligent Environments". *In Proc. IEEE Workshop on Human Modelling, Analysis and Synthesis*, 2000.

[23] M. Villamizar, J. Scandaliaris, A. Sanfeliu, and J. Andrade-Cetto. Combining color invariant gradient detector with HOG descriptors for robust image detection in scenes under cast shadows. In *Proceedings of the International Conference on Robotics and Automation, ICRA*, 2009.

[24] P. Viola and M. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57:137–154, 2004.