Data Fusion, Confidence Assessment and Intelligent Sensor Recruitment in Multimodal Tracking

Zenon Mathews[1], Sergi Bermúdez i Badia[1], Ulysses Bernardet[1], and Paul F.M.J. Verschure[1]

Universitat Pompeu Fabra, Barcelona

Abstract. In mixed reality environments many applications require reliable tracking of movements of objects and people. Often many different sensors with different characteristics are deployed to tackle this problem. We show how to fuse multimodal information to improve tracking and we propose a method for intelligent recruitment of sensors to resolve conflicting situations. We show that probability maps for sensors can be automatically generated to assess the credibility of sensors in different regions of the tracked area. Tests are conducted using wall-mounted and overhead cameras, gazers and a pressure sensitive floor.

1 Introduction

Our project is in the framework of a mixed-reality environment called *P-club* which is being developed to facilitate the understanding and exploitation of brain mechanisms for *presence* [2]. For this purpose it is indispensable to have a reliable tracking system, which is capable of supplying the information about the movements of real subjects in the physical space of *P-Club*. The interactive space *Ada* is used as the basis for constructing this mixed-reality space [6]. *Ada* can be considered to be an artificial organism in the shape of an environment, an inside-out robot, that has its own goals and expresses its own internal states. *Ada* comprises a pressure sensitive floor, pan-tilt cameras, light-fingers, triples of microphones for sound recognition and localization, a 360° projection screen and 14-channel sound output. This sensor-effector system is controlled by a neuromorphic system simulated on a cluster of about 30 PCs. *Ada* was exhibited to the public for 5 months in Switzerland during 2002 and visited by about 550.000 people (figure 1 left panel).

Different sensors have data of different quality and nature and should be fused together. Additionally, a confidence measure of each of the sensors is convenient during the process of fusion. Further, from time to time the sensor data would be conflicting or inadequate. We propose methods for data fusion, automatic assessment of sensor data qualities and intelligent deployment of sensors to gather more information about objects in case of conflicting or scanty data. We present experimental results of our methods and an outlook for the multimodal tracking project.

2 Methods

In this project we use on one hand wall-mounted cameras and gazers and on the other hand a pressure sensitive floor for the multimodal input. Further sensors can be easily added to the existing framework to enhance the multimodality, e.g. auditory localization systems. Sensor data-exchange for data-fusion is fairly straightforward, as is implemented using a TCP/IP socket for communication, on the condition that each sensor-system complies with a predefined protocol. This ensures a common interface independent of the sensor type or quality. The visual tracking and the floor sensory-systems deliver presence of objects as two dimensional Gaussian distributions with a mean (i.e. estimated x, y coordinates) and standard deviations, whereas the gazer-system gives color-histograms of objects as data.

2.1 Input Systems

Overhead Tracking We use the custom built tracking system called AnTS to track moving objects using overhead cameras [3]. This system delivers position estimation of objects as two dimensional Gaussian distributions. The software AnTS and all the tools contained in it where developed in C++ under Linux. The core of the system is the image processing part, and for it Intels Open Source Computer Vision Library (OpenCV), a powerful multiplatform and open source image processing library, was used [5].

Pressure Sensitive Floor The pressure sensitive floor tiles were already used in the interactive space called *Ada* for Expo 2002 in Switzerland (figure 1 left panel). The floor is a sensor and at the same time an effector as each of its tiles can display different colours (figure 1). In order to avoid the interference of floor-illumination with the visual tracking and for reasons of controllability and scalability we simulated the floor in our preliminary experiments. This floor simulation is driven by an overhead camera. The simulation was implemented using Java3D Library for Java and simulates movements of people/objects on the hexagonal floor tiles of *Ada* (figure 1 left panel). This simulation will in the near future be replaced by the real floor infrastructure.

A floor tile has a hexagonal shape and can either be 'on' or 'off'. The pressure sensitive floor tiles are switched 'on' when an object or person exerts pressure (i.e. stays on it). The floor is simulated in the following way: the input from the overhead camera tracking system is mapped onto the floor by setting the closest floor tile to 'on'. For that we measure the distance between the position estimation from the camera and the midpoint of the floor tiles and select the closest floor tile (1 right panel).

Gazers The gazers, movable color cameras with controllable zoom, are controlled via a custom-built DMX-interface and will be used to capture images of objects from which we extract color histograms [4]. We will propose a method

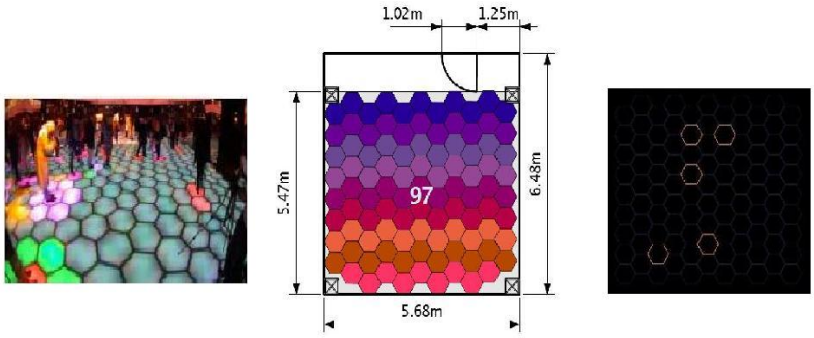

Fig. 1. *Ada* floor layout and the floor simulation. **Left:**Pressure Sensitive floor of Ada in operation at the Expo 2002 CH (adapted from [9]).**Middle:** Layout of the pressure sensitive floor infrastructure. **Right:** Screenshot of a running floor simulation: the overhead camera input is mapped onto the nearest floor tile. Lighted-up floor tiles represent tracked objects.

to deploy them intelligently to gather information about objects in case of conflicting data from the other two sensor systems mentioned before. This proposed method will be extended to involve more sensors and effectors in the future.

2.2 Data Fusion

The above mentioned systems, and also further ones to be augmented in future, are assumed to deliver Gaussian distributions as position estimations for detected objects. We use a Gaussian distribution merging technique described in [1] and [7] to fuse the data from different sensors.

Merging Gaussian Distributions We differentiate between local L and global G coordinate frames. Each sensor system acts in its own local frame and this should be transformed into the global frame used by the $Fuser$. Let C_{L_i} be the covariance matrix of sensor number i. This matrix can be determined from the major and minor distribution axis standard deviations of the local coordinate frame of the sensor i.

$$C_{L_i} = \begin{pmatrix} \sigma^2_{maj} & 0 \\ 0 & \sigma^2_{min} \end{pmatrix} \tag{1}$$

where σ^2_{maj} and σ^2_{min} are the major and minor axis standard deviations. And we assume that the observations of sensor i are oriented at an angle θ with respect to the global x-axis. Therefore we need to perform the following rotation of the covariance matrix.

$$C_{L_i} = R(-\theta)^T C_{L_i} R(-\theta) \tag{2}$$

After this canonisation we can combine the individual covariance matrices C_{L_i} and C_{L_j} into one representing the combined distribution (C_{L_j} being the covariance matrix of another sensor system)

$$C^{'} = C_{L_i} - C_{L_i}[C_{L_i} + C_{L_j}]^{-1}C_{L_i} \tag{3}$$

And the mean X of the merged distribution is computed from the individual means $\hat{X}_i$ and $\hat{X}_j$ and the individual covariances.

$$\hat{X}^{'} = \hat{X}_i + C_{L_i}[C_{L_i} + C_{L_j}]^{-1}(\hat{X}_j - \hat{X}_i) \tag{4}$$

Moreover, the principal axis angle can be calculated from the merged covariance matrix entries.

$$\theta^{'} = \frac{1}{2}tan^{-1}\left(\frac{2C'_{12}}{C'_{11} - C'_{22}}\right) \tag{5}$$

And finally we can compute the major and minor axis standard deviations by rotating the covariance matrix to align the axes.

$$C^{'} = R(\theta^{'})^T C' R(\theta^{'}) \tag{6}$$

Construction of a Credibility Map While merging the data from different sensors we would like to automatically build a confidence map for each sensor, enabling confidence estimations for sensor readings in relation to the object position. For this we estimate the relative error of sensor i, given the sensor reading (x_i, y_i) for a specific tracked object and the merged result (x', y'). Then the relative error of sensor i is estimated as:

$$e_i = \frac{|x_i - x'| + |y_i - y'|}{w + h} \tag{7}$$

where w and h are the width and height of the tracked area. For each sensor measurement and fused result pair, we can calculate this relative error. With n such pairs we calculate the confidence map of our sensors through a two dimensional interpolation.

Such confidence maps reflect how reliable each sensor is in the different regions of the tracked area. For example an overhead camera might have perspective distortions, making its readings more erroneous along the edges of its view than in its interior.

2.3 Intelligent Recruitment of Sensors

Tracking systems often run into conflict situations, e.g. when trajectories of tracked objects intersect, where there are no means of regaining object-IDs from the fused data. The solution is to actively gather further data about the characteristics of the objects. We propose a *Smart-Sensor-Recruiter* (further called *Recruiter*) for dynamically deploying sensors to detect and resolve these conflicting situations (figure 2). The confidence map generated for each sensor-system

can be used by the *Recruiter* to recruit the best sensors for the given position in the tracked area. The *Fuser* communicates the object positions and their standard deviations to the *Recruiter*. The *Recruiter* takes actions to predict and resolve conflicting situations and supplies the *Fuser* with appropriate information, such as color histograms etc.. This is done by dynamically deploying sensors to gather specific information about objects or influencing the sensor-readings by changing the tracking condition (floor-illumination, light-fingers etc.).

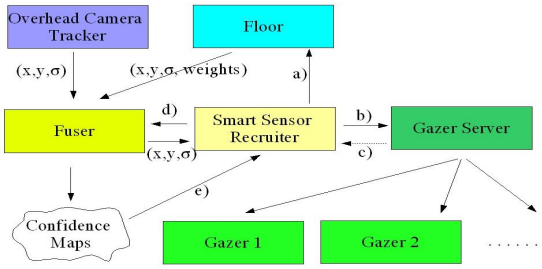

Fig. 2. Smart deployment of sensors for conflict resolution. Only the *Fuser* maintains IDs of tracked objects. It regularly feeds the coordinates and std. deviations of objects to the *Recruiter*. The *Recruiter* predicts conflicts and supplies adequate information for its resolution to the *Fuser*. It thereby makes use of the confidence-map and actively recruits sensors, e.g. the gazer-server (arrow *b*), which can supply color-histograms of tracked objects (arrow *c*). And the floor can be recruited (arrow *a*) to light all its tiles to white to improve the performance of visual tracking systems. Analyzing the color-histograms from the *Recruiter*, the *Fuser* can maintain object-IDs. The *Recruiter* can be made to learn such sensor-recruiting tasks using adaptive systems like DAC5 [8].

The *Recruiter* has the task of deciding when and what actions to take to supply the *Fuser* with enough information about objects to guarantee correct object-IDs at any given time. For example the *Recruiter* could look at the trajectories of objects and assess the risk of a potential collision situation. If they get too close and are on a collision course the *Recruiter* employs movable gazers to extract color histograms of the objects. After a path intersection, the *Recruiter* triggers the information retrieval again and the *Fuser* compares theses information (e.g. color histograms) with the ones before the objects met and regains the object-IDs.

Such deployment of sensors is in no way restricted to gazers. Any sensor capable of supplying some characteristic information about the tracked objects can be employed by the *Recruiter*. For example, an auditory tracking system possessing sound localization techniques. Thereby the *Recruiter* can use the automatically generated confidence maps to decide which sensor to recruit. De-

sirably the *Recruiter* could learn how and when to recruit and deploy particular sensors using adaptive learning systems like DAC5 [8].

DAC5 is a general purpose adaptive learning model, which uses three strongly coupled control layers, *reactive*, *adaptive* and *contextual*, to learn the appropriate reactions to stimuli. Using robot based models of DAC5, it was shown that those robots could learn complex rules in real-world situations.

3 Experimental Results

We tested the methods described above using a wall-mounted camera for driving the simulated floor, an overhead tracking system and gazers for histogram extraction. Inter-process communications between the individual sensors and the data-fuser and floor-simulation were realized as TCP/IP sockets.

We used Gaussian fusion method to improve the accuracy of the overhead tracking by combining it with the floor data. The standard deviations of the sensor data describes the confidence of the measurement it provides. Given a higher standard deviation, a more uncertain measurement is expected. Then, we show how multimodal fusion significantly reduces the uncertainty (figure 3). The data collected in the experiment is used to generate the confidence map for the camera sensor (figure 4). This map clearly shows that the error of the overhead camera is higher in the more distorted areas. This information is then used by both the *Fuser* and the *Recruiter*. Moreover each individual object provides a unique color distribution that can be used for their identification (figure 5).

The images from gazers triggered by *Recruiter* can be used to extract colour histograms of different tracked objects (5). This helps the *Fuser* in object-ID resolution.

4 Conclusion and Outlook

We propose a multimodal infrastructure to solve the tracking problem in a mixed reality environment. In this context we show how multimodal information can be fused to gain accuracy and confidence in our measurement. Subsequently, we propose a framework for automatic generation of confidence maps for sensors in a multimodal setup and discuss how this can be used to intelligently deploy sensors for conflict resolution.

In the further development of this project we would like to extend the system to be able to dynamically incorporate newly available sensors. A publish-and-subscribe method is planned, where available sensors can publish themselves and the *Fuser* can subscribe them. Although we used a Gaussian merging technique we plan to employ more general statistical data fusion methods like Bayesian inference. Once a confidence map has been created for each sensor, this a priori information will be used for the Bayesian inference. Another major step is the implementation of an adaptive learning mechanism like DAC5 for the *Recruiter*, that will learn how to intelligently deploy sensors and actively modify the environment to provide the best tracking condition.

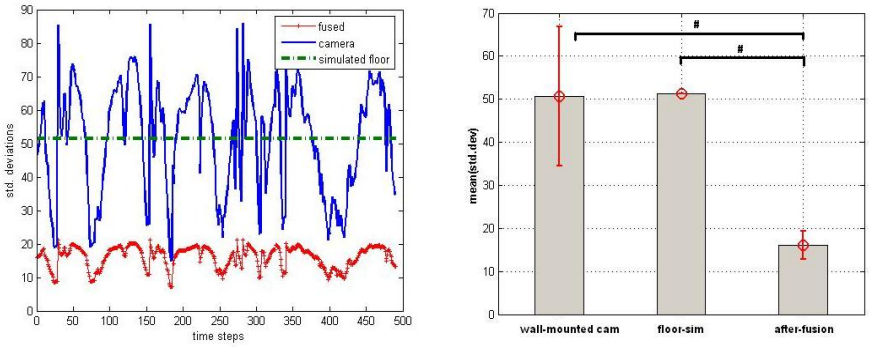

Fig. 3. Certainty improvement through data fusion.**Left:** Standard deviations for the merged and the individual sensor data. We used the visual tracking (figure 4) and the simulated floor (driven by another wall-mounted camera) as two different sensors. In the simulation we set the floor to always have the same standard deviation respecting the equivalence of all floor tiles. We obtain a more accurate estimation of position in the merged distribution. **Right:** The mean of std. deviations. A T-test at significance level $p = 0.05$ shows that the fused certainty is significantly higher than the individual ones (#).

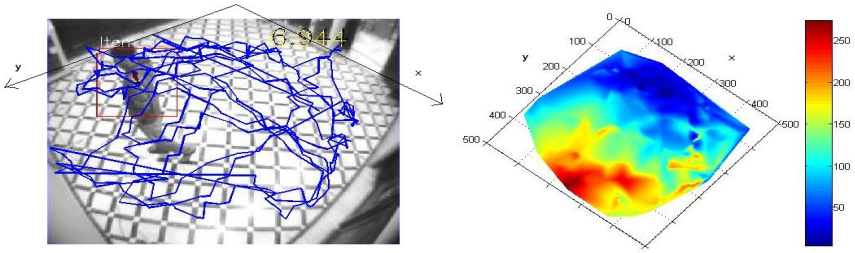

Fig. 4. Confidence map for visual tracking.**Left:** Fused object trajectory projected onto the view from the overhead camera. **Right:** The confidence map ((x, y) against the relative error of the sensor as described earlier) for the panel on the left, using cubic interpolation after data-fusion. The measurements are more erroneous the bigger the lens distortion and this is captured in the confidence map. Such confidence maps can be built for other sensors to dynamically assess the data quality in relation to the tracked area, capacitating adaption to dynamic changes of sensor behaviors.

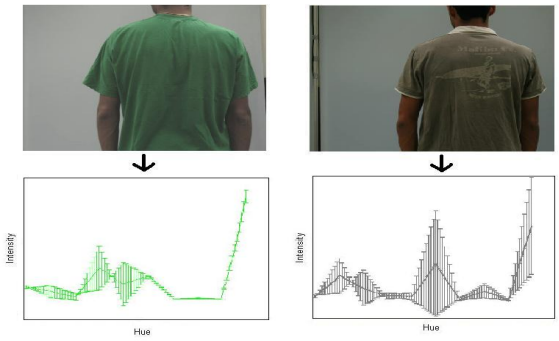

Fig. 5. Object recognition via color histogram analysis. Images of tracked objects with various colours (top) have different colour histograms (bottom). The plots on the bottom panel show hue-intensities and standard deviations from many snapshots of the same person. A Kolomogorov-Smirnov-test gives the following results: $D = 0.2198$ with a corresponding $P = 0.021$. Analysis of such color histograms generated from snapshots by gazers can help the *Fuser* to differentiate or identify tracked objects.

Acknowledgments. This project is supported by the European PRESENCCIA (IST-2006-27731) project.

References

1. Stroupe A. W., Martin C. M., Balch T., *Merging Gaussian Distributions for Object Localization in Multi-Robot Systems*, D. Rus and S. Singh (Eds.):Experimental Robotics 7, LNCIS 271, pp. 343-352, 2001 © Springer-Verlag Berlin.
2. Zimmerli L., Duff A., Mura A., Eng K., Bermúdez i Badia S., Bernardet U., Mathews Z., Verschure P.F.M.J.: *Communication and Interaction in the Persistent Mixed Reality Environment P-Club.* Interactive media, The enhancement of multilingual communication and learning through technology, Ascona III November 2006, Ascona, Switzerland.
3. Bermúdez i Badia S.: *The Principles of Insect Navigation applied to Flying and Roving Robots: from Vision to Olfaction.* Dissertation submitted to the Swiss Federal Institute of Technology Zürich 2006.
4. Lovis P.: *Hello Stranger DMX Interface and Simple DMX Framework.* Semester Thesis October 2005, Institute of Neuroinformatics, ETH Zürich.
5. OpenCV: *Intel Open Source Computer Vision Library.* http://www.intel.com/research/mrl/research/opencv.
6. Eng K.: *Designing neuromorphic interactive spaces.* Dissertation submitted to the Swiss Federal Institute of Technology Zürich 2004.
7. Anderson W. and Duffin R.: *J Mathematical Analysis Applications 26:576, 1969.*
8. Verschure P.F.M.J., Althaus P.: *A real-world rational agent: unifying old and new AI.* Cognitive Science 27 (2003) 561..590.
9. Delbrück T., Whatley A.M., Douglas R., Hepp K.: *The mother of all disco floors, Institute of Neuroinformatics.* White Paper, ETH/Univ. Zürich, Switzerland.