This document is published in:

Andre Ponce de Leon F. de Carvalho, et al. (eds.) (2010). *Distributed Computing and Artificial Intelligence: 7th International Symposium*. (Advances in Intelligent and Soft Computing, 79) Springer, 301-308.
DOI: http://dx.doi.org/10.1007/978-3-642-14883-5_39

Multi-camera and Multi-modal Sensor Fusion, an Architecture Overview

Alvaro Luis Bustamante, José M. Molina, and Miguel A. Patricio

Applied Artificial Intelligence Group, Universidad Carlos III de Madrid, Avd. de la Universidad Carlos III, 22, 28270, Colmenarejo, Madrid, Spain e-mail: {alvaro.luis,miguelangel.patricio}@uc3m.es, josemanuel.molina@uc3m.es http://www.giaa.inf.uc3m.es

Abstract. This paper outlines an architecture for multi-camera and multi-modal sensor fusion. We define a high-level architecture in which image sensors like standard color, thermal, and time of flight cameras can be fused with high accuracy location systems based on UWB, Wifi, Bluetooth or RFID technologies. This architecture is specially well-suited for indoor environments, where such heterogeneous sensors usually coexists. The main advantage of such a system is that a combined non-redundant output is provided for all the detected targets. The fused output includes in its simplest form the location of each target, including additional features depending of the sensors involved in the target detection, e.g., location plus thermal information. This way, a surveillance or context-aware system obtains more accurate and complete information than only using one kind of technology.

1 Introduction

Video surveillance has been the most popular security tool for years. Banks, retail stores, and countless other end-users depend on the protection provided by video surveillance. And thanks to the new breakthroughs in this evolving technology, security cameras are more effective, cheaper, and easy to deploy than even before.

This advances has issued the increase of image sensors, thanks in part also to the IP-based video emerging technology, opening a new research field in the last decade. The huge amount of video sensors installed in some scenarios makes unaffordable use humans operators for real-time monitoring. This way, new automated tracking systems are proposed in order to solve this problem [13]. These systems addresses the task of multiple people tracking in multi-camera environments. So, many effort put in this area consists in perform fusion information provided by the different

optical sensors, solving problems such as background and foreground detection, object tracking, tracking occlusion, track continuity, and so on [18].

Meanwhile, with the technical advances in ubiquitous computing [1], wireless networking and the proliferation of mobile computing devices, there has been an increasing need to capture the context information for context-aware systems and services. Context-aware computing is a mobile computing paradigm in which applications can discover and take advantage of contextual information (such as user location, time of day, nearby people and devices, and user activity) [3]. Therefore, much research has focused on developing services architectures for location-aware systems [16], and also many attention has been paid to the fundamental and challenging problem of locating and tracking mobile users, especially in in-building environments, since, as discussed in [9], context-aware systems are based fundamentally in the user location. Hence, new systems for indoor location have emerged using different wireless technologies such as Wifi [14], Ultra Wide Band (UWB) [5], Radio Frequency IDentification (RFID) [6], etc.

Such deploy of multi-camera environments, indoor-localization systems, and automated specific processes and services, has allowed the coexistence of both video surveillance and indoor location systems in the same environment, but usually with different scopes. Video surveillance and automated tracking systems are fundamentally used for security purposes such as intrusion and event detection, activity recognition, or simply as *a posteriori* forensic tool. Meantime, indoor location is used especially for context-aware services, access control, personnel monitoring, augmented reality, etc.

The differenced use of both kind of technologies coexisting in the same environment is feasible, but could be improved if both techniques complements each other. For example, consider a location platform which provides a rich information of each target, such as, high-accuracy location, trajectory, speed, shape, color, size, thermal information, real-time video, and so on. One single sensor cannot provide all this information, so a fusion platform is needed to process all the independent sensor data and provide a single fused source of information. All this features could be used in complex systems like event recognition [12], behavioural profiling [2], action recognition [17], or simply, advanced surveillance and context-aware systems.

This way, our proposal consists in a hybrid fusion architecture which enables the fusion information of image and location sensors. Fuse these different sources is not a trivial task, therefore we define a first approach clarifying the different aspects, processes, and design decisions involved in such fusion architecture. In this topic there are not a sizeable literature, since most effort in the fusion field has been placed in fuse sensors of similar characteristics [8]. Some works, as described in [4, 11] deals with vision and location sensors fusion, but appears to be *ad-hoc* solutions to specific problems.

The rest of the paper is organized as follows: Section 2 overviews the fusion architecture, defining their basic inputs and outputs. Section 3 describes in more detail the architecture, paying special attention to the more relevant parts of the system.

Section 4 concludes with some reflexions about the architecture and describing the future work.

2 Architecture Overview

The main idea of the architecture is to be able to process multiple location and image sensors and provide a single fused, non-redundant output. The sensors for location could be Wifi, UWB, RFID, Bluetooth, etc. In fact, any wireless technology which can provide a location of a target with their associated identifier (i.e, the MAC address of the location device). In the vision field can be used standard image color sensors, thermal cameras, infrared sensors, time of flight cameras, and so on. Figure 1 represents the high-level input and output of the desired architecture.

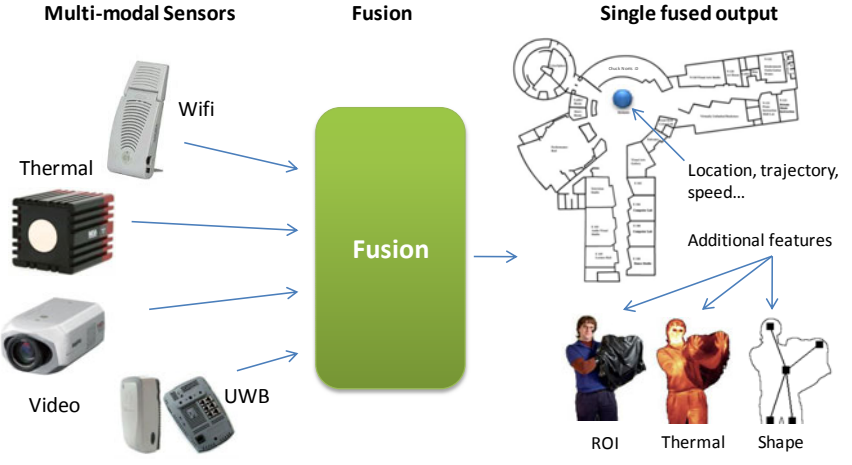

Fig. 1 High-level fusion architecture input/output

The architecture should provide all the information available of each target for each client subscribed to the fusion system, not only the fused location. That is, if a fusion is performed between a UWB location system and a color image sensor, the system should provide a single fused location and also the additional information offered by a image sensor, like color, shape, the image of the target, etc. So *a posteriori* processing could be done if required.

Regarding the final users or systems accessing the fusion architecture, the goal is that final output were accessible from any subscribed client, and therefore, be accessible both by surveillance and context aware systems at the same time. So, in some way the fusion system should also acts as a location server.

The full specification of the proposed architecture, with all their algorithms, protocols, etc, would exceed the length of the paper. Instead, we provide a brief overview of the different parts.

3 Detailed Fusion Architecture

Processing algorithms can be organized in different fusion architectures. The proposed solution is organized in different distributed tiers, each one processing inputs from the bottom tiers and feeding the upper one. The bottom tier is associated with the local sensor processing, so each sensor is the responsible of process it own information and provide a list of local tracks. This is generally achieved by the local track processing module as shown in figure 2.

Depending on the set of sensors used in the system, may be required an intermediate fusion step. As shown in figure 2, color and thermal image sensors are previously fused before the general location fusion. This particular sensor fusion node can achieve advantage in the fusion step, since sensors of the same type can obtain similar target features in order to improve fusion, i.e., in the color sensor fusion node, we can use attributes like location, shape, size, color, etc, to perform a better fusion than only fusing tracks locations. Anyway, this previous step is only necessary when there are some sensors of the same type with overlapped vision.

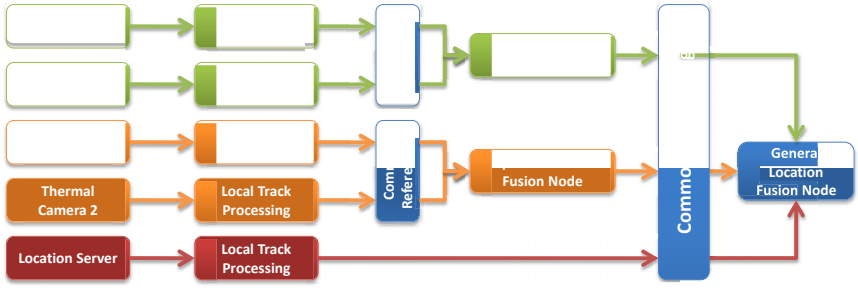

Fig. 2 Detailed fusion architecture

The second tier fuses all the local tracks provided by each sensor or set of sensors and generates a set of non-redundant global tracks, as the output of the system. This fusion step only use location to combine the local tracks, due to this is a common attribute in the local tracks provided by underlying tiers in a heterogeneous sensor network.

The main advantage of this architecture resides in the distributed data processing which lets each processing adjusts to the particularities of each sensor. This advantage is highlighted in this particular system, where imaging and location sensors are used. In this way particular processing algorithms could be defined for each sensor type. Another advantage of a decentralized architecture could be that the computational load could be balanced across different processors executing each one of the tasks on the system.

Each system module and some considerations are explained in more detail in the following sections, overviewing the different aspects needed to be taking into account when developing such fusion architecture.

3.1 Local Track Processing

The local track processing is the part of the architecture which deals directly with the data streams provided by sensors. The input to this module is the data provided by their associated sensor (an image in image sensors and locations in location sensors). In their turn, the output should be a set of local tracks, that is, all the targets detected by the sensor, with all their available features. This system should keep updated the track list in various ways, i.e, updating existing tracks with new associated plots, creating new tracks, and deleting tracks after some lack of updates. This module can be outlined in figure 3 where a simple local track processor is presented. In such processor, classical gating, association and filtering processes are performed [8].

Image sensors addresses the problem that cannot provide a location (plots) of the moving targets present in the scene, as they only provide an image. So is needed in this case apply a real-time object tracking based on image analysis [18]. In the other hand, location sensors usually provides some kind of identification with each target location, so normally is not needed a preprocessing step.

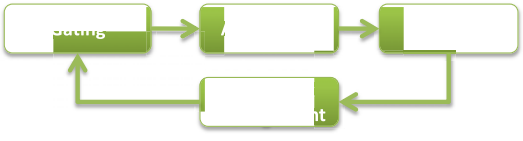

Fig. 3 Local sensor processing overview

3.2 Common Referencing

Location-based fusion using sensors of diversity nature also introduces a handicap when trying to represent all locations in a common coordinate system. Usually, camera tracking is achieved directly over the image, that is, in the camera perspective of the scene, and this is a 2D representation with X and Y pixels coordinates. Location sensors usually lets the user establish the coordinate system and its location, so in this case the main problem arise with image sensors.

Fortunately, there are many approaches in the multi-camera fusion literature in order to provide a common referencing between multiple views. The most used is the based in the concept of homography [10]. In the computer vision field, any two images of the same planar surface in space are related by a homography (assuming a pinhole camera model). This has many practical applications, such as image rectification or image registration. For example is possible to change the perspective view of a camera and then process a synthesized image plane, as shown in figure 4.

In any case, this module must be able to transform the location and speed of each local track, in a common coordinate system. This way, fusion nodes can perform, at least, the location-based fusion.

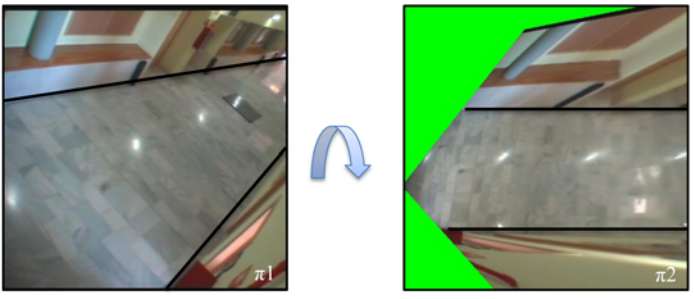

Fig. 4 Projective plane transformation

3.3 Fusion Nodes

A single fusion node is the responsible of fuse all the input tracks provided by each local track processor or another fusion node, and provide a single fused, non-redundant, set of tracks. In figure 2 we define two different types of fusion nodes, one sensor-specific, and other more general based only in location. This differentiation is due to the sensor-specific implementation must take into account the extra features provided by a set of common image sensors, like, color, shape, temperature, etc, achieving a location and feature-based fusion. The feature-based fusion has been widely used and tested in many fields with successful [7], so is useful distinguish both kind of fusion nodes. The general location-based fusion achieved in the second level only should consider the location of the tracks to perform the fusion, since there are not common features between, i.e., a image sensor and a location sensor that provide X, Y, Z coordinates.

The fusion nodes, are also the responsible, as the local track processors, of maintain updated the set of output tracks, attending to the input tracks. So, it should manage the creation, update and deletion when required. Classical processes of gating, association and filtering are also performed over the tracks, as described in the local track processors.

Would be also useful that fusion nodes could provide for each output track, the input tracks identifiers that has contributed to generate them. This way, the client regarding the output of the second level fusion node could know all the local tracks contributing for a final global track, and then, know all the independent features of a global track.

3.4 Infrastructure Considerations

There is an inherent problem in the fusion architecture proposed, and is the transmission of all the information from video and location sensors over the different local track processors, fusion nodes, and the final client when required. The decentralized architecture enables a distributed sensor processing, so is needed to enable some infrastructure allowing multiple video and general data transmission. For both image and location sensors is required to be enabled independent servers

which could broadcast all the information over the same network. For image sensors we propose the digital video streaming system described in [15], implementing a JPEG2000 RTP streaming service allowing broadcast transmissions. Some similar location server must be enabled for each location system attached to the architecture.

4 Conclusions and Future Work

In this paper, a overview of an architecture for multi-camera and multi-modal sensor fusion has been given. The architecture is scalable in the way that many heterogeneous image and location sensors can be attached to the system. This provides an improved location service, taking the advantages of the different sensors used. The architecture is the responsible of processing efficiently all the data sensors, so a distributed sensor processing is proposed, with all their inherently benefits.

We have noticed along the description of this architecture that such system can be very complex to design, so it is needed to define well all the aspects and algorithms involved. In future works the different processing algorithms, communication protocols, and other important aspects of the architecture will be further described. Working prototype of the architecture is being developed using the infrastructure of VISLAB, with some color image sensors, one thermal camera, and a UWB indoor localization system.

Acknowledgements. This work was supported in part by Projects CICYT TIN2008-06742-C02-02/TSI, CICYT TEC2008-06732-C02-02/TEC, SINPROB, CAM CONTEXTS S2009/TIC-1485 and DPS2008-07029-C02-02

References

1. Abowd, G., Mynatt, E.: Charting past, present, and future research in ubiquitous computing. ACM Transactions on Computer-Human Interaction (TOCHI) 7(1), 58 (2000)
2. Atallah, L., ElHelw, M., Pansiot, J., Stoyanov, D., Wang, L., Lo, B., Yang, G.: Behaviour profiling with ambient and wearable sensing. In: 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007), pp. 133–138. Springer, Heidelberg (2007)
3. Chen, G., Kotz, D.: A survey of context-aware mobile computing research. Technical report, Citeseer (2000)
4. Germa, T., Lerasle, F., Ouadah, N., Cadenat, V.: Vision and RFID data fusion for tracking people in crowds by a mobile robot. Computer Vision and Image Understanding (2010)
5. Gezici, S., Tian, Z., Giannakis, G., Kobayashi, H., Molisch, A., Poor, H., Sahinoglu, Z.: Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks. IEEE Signal Processing Magazine 22(4), 70–84 (2005)
6. Hahnel, D., Burgard, W., Fox, D., Fishkin, K., Philipose, M.: Mapping and localization with RFID technology. In: Proceedings of IEEE International Conference on Robotics and Automation, 2004. ICRA 2004, vol. 1 (2004)
7. Hall, D., Llinas, J.: An introduction to multisensor data fusion. Proceedings of the IEEE 85(1), 6–23 (1997)
8. Hall, D., Llinas, J.: Handbook of multisensor data fusion. CRC, Boca Raton (2001)

9. Harter, A., Hopper, A., Steggles, P., Ward, A., Webster, P.: The anatomy of a context-aware application. Wireless Networks 8(2), 187–197 (2002)
10. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge Univ. Pr., Cambridge (2003)
11. Hofmann, U., Rieder, A., Dickmanns, E.: Radar and vision data fusion for hybrid adaptive cruise control on highways. Machine Vision and Applications 14(1), 42–49 (2003)
12. Hongeng, S., Nevatia, R., Bremond, F.: Video-based event recognition: activity representation and probabilistic recognition methods. Computer Vision and Image Understanding 96(2), 129–162 (2004)
13. Kanade, T., Collins, R., Lipton, A., Burt, P., Wixson, L.: Advances in cooperative multi-sensor video surveillance. In: Proceedings of DARPA Image Understanding Workshop, Citeseer, vol. 1, p. 2 (1998)
14. Lim, H., Kung, L., Hou, J., Luo, H.: Zero-configuration, robust indoor localization: theory and experimentation. In: Proceedings of IEEE Infocom, Citeseer, pp. 123–125 (2006)
15. Luis, A., Patricio, M.: Scalable Streaming of JPEG 2000 Live Video Using RTP over UDP. In: International Symposium on Distributed Computing and Artificial Intelligence 2008 (DCAI 2008), pp. 574–581. Springer, Heidelberg (2008)
16. Maass, H.: Location-aware mobile applications based on directory services. Mobile Networks and Applications 3(2), 157–173 (1998)
17. Yamato, J., Ohya, J., Ishii, K.: Recognizing human action in time-sequential images using hidden Markov model. In: Proc. Comp. Vis. and Pattern Rec., pp. 379–385 (1992)
18. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Computing Surveys (CSUR) 38(4), 13 (2006)