A Novel Framework for Data Registration and Data Fusion in Presence of Multi-modal Sensors

Hadi Aliakbarpour[1,2], Joao Filipe Ferreira[1], Kamrad Khoshhal[1], and Jorge Dias[1]

[1] Institute of Systems and Robotics (ISR), University of Coimbra, Portugal
{hadi,jfilipe,kamrad,jorge}@isr.uc.pt
[2] IEEE student member

Abstract. This article presents a novel framework to register and fuse heterogeneous sensory data. Our approach is based on geometrically registration of sensory data onto a set of virtual parallel planes and then applying an occupancy grid for each layer. This framework is useful in surveillance applications in presence of multi-modal sensors and can be used specially in tracking and human behavior understanding areas. The multi-modal sensors set in this work comprises of some cameras, inertial measurement sensors (IMU), laser range finders (LRF) and a binaural sensing system. For registering data from each one of these sensors an individual approach is proposed. After registering multimodal sensory data on various geometrically parallel planes, a two-dimensional occupancy grid (as a layer) is applied for each plane.

Keywords: Multi-modality, data registration, data fusion, Occupancy grid and Homography.

1 Introduction

Data registration and *data fusion* are two crucial issues in a multi sensory environment since heterogeneous sensors are substantially different. In this work we propose a novel framework for data registration and fusion in presence of various modalities such as image, range, sound and inertial data. This paper is structured as following. In Sec. 2 the main contribution is described. Sec. 3 is for related work. Sensors models are introduced in Sec.4. Sec. 5 is dedicated to registering data from different modalities and then applying a data fusion strategy. A preliminary result of an ongoing experiment is shown in Sec. 6. Then in Sec. 7 conclusion and future work are discussed.

2 Contribution to Technological Innovation

The use of surveillance systems is growing in different areas such as airports, banks and human behavior interpretation. Each particular sensor has its own drawbacks. Although to overcome that nowadays researchers are trying to use several sensors from different modalities instead of using just a single type, however data registration and data fusion are still two critical and determinant phases in surveillance systems. Our main contribution is to work on these subjects and make a framework based on geometric data registration and probabilistic data fusion.

L.M. Camarinha-Matos, P. Pereira, and L. Ribeiro (Eds.): DoCEIS 2010, IFIP AICT 314, pp. 308–315, 2010.

3 Related Works

Fusion of the sensors outputs is a crucial challenge related in such applications. A model namely JDL is proposed in [1] for data fusion. Armesto et al. in [2] presented an approach to model and fuse of non-linear data from multi-rate systems which have visual and inertial sensors. Bellotto and Hu also presented an approach in [3] to fuse the LRF data and vision data by PTZ camera. Chakravarty and Jarvis in [4] discussed a fusion method for LRF and panoramic vision data to track people from a stationary robot simultaneously. Khan et al. in [5] presented a new approach which fuse different views silhouette data form some.

4 Sensors Models

A pinhole camera model [6] is used in this work. The 3D LRF in this work is built by moving a 2D LRF along one of its axes. A detailed model of this configuration can be found in [7]. A detailed version of IMU model used in this paper can be seen in [12]. And eventually for the sound system, a Bayesian binaural system is considered for imitating human binaural system (see [8]). The Bayesian binaural system (described in [9]) composes three distinct and consecutive processors (Fig. 3-a): *the monaural cochlear unit*, which processes the pair of monaural signals $\{X_1,X_2\}$ coming from the binaural audio transducer system by simulating the human cochlea, so as to achieve a tonotopic representation (i.e. a frequency band decomposition) of the left and right audio streams; *the binaural unit*, which correlates these signals and consequently estimates the binaural cues and segments each sound-source; and, finally, *the Bayesian 3D sound-source localization unit*, which applies a Bayesian sensor model so as to perform localization of sound-sources in 3D space, using the egocentric frame of reference $\{\varepsilon\}$.

5 Multi-modal Data Registration and Data Fusion

In a multi-modal sensory environment, data registration and data fusion are two crucial issues. In this section we talk about these issues where there are four heterogeneous sensors. Prior to that, a coordinate reference $\{W\}$ and also a 3D plane π_{ref} inside this space will be introduced which are universal and common for all sensors inside the framework (see fig. 1-left). Let the origin of this coordinate frame be on our reference plane and also consider two coordinate vectors X and Y extending across (aligned) the plane (XY-plane of $\{W\}$ corresponds to π_{ref}). Moreover assume that the Z direction of this plane is parallel to the earth gravity vector but in an apposite direction. Having these definitions, π_{ref} becomes *a virtual horizontal plane* [10] which is considered a common geometrical plane for all sensors in this setup. The idea is to geometrically register data observed by sensors with different modalities (cameras, LRFs and microphones) onto this plane (in section 5.4 it is extended to some parallel planes). After having the registration structure, a BOF [14] will be applied to achieve the final framework.

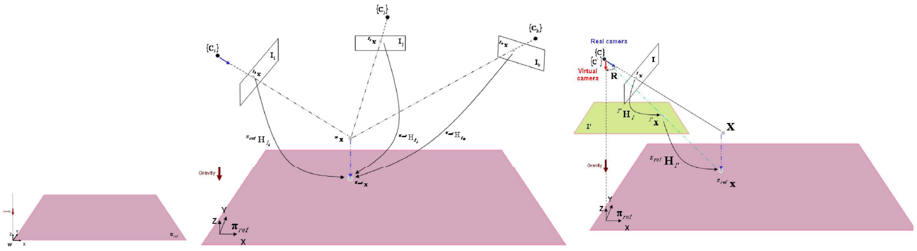

Fig. 1. ***Left***: World coordinate system, W, and the reference plane, π_{ref}. ***Middle***: Registering image points onto the reference plane π_{ref}. ***Right***: Considering a virtual camera for each real camera. The principle axis of such a virtual camera is coincided to the gravity vector. The rotation matrix between real camera $\{C\}$ and virtual camera $\{C'\}$ is given by the IMU [12].

5.1 Image Data Registration

The idea of this work is to have all data coming from different sensors registered on a common reference plane, namely π_{ref}. Fig. 1-*middle* shows a setup with n_c cameras, $C = \{C_i \mid i = 1..n_c\}$, and a reference plane π_{ref}. In this setup each camera is rigidly fixed with an IMU (Inertial Measurement Unit). The intention is to register a 3D point ${}^{W}X = (X,Y,Z)$ observed by camera C onto the reference plane π_{ref} as ${}^{\pi_{ref}}x$ using the IMU. The image plane of these cameras can be expressed as $I = \{I_i \mid i = 1..n_c\}$. A virtual image plane, $\{I_i' \mid i = 1...n_c\}$, is considered for each camera. Such a virtual plane is assumed to be a horizontal plane at a distance f below the camera sensor, f being the focal length [10]. In other words, it can be said that beside of each real camera C in the setup, a virtual camera *C'* is also considered whose center, {C'}, is coincided to the center of the real camera {C}. So that the transformation matrix between {C'} and {C} will just have a rotation part and the translation part is a zero vector. In order to have ${}^{\pi_{ref}}x$ from ${}^{W}X$ three consecutive steps are considered: Firstly, a 3D point ${}^{W}X$ is projected on the camera reference frame ({C}) using ${}^{I}x = P\,{}^{W}X$ (*P* is projection matrix). Secondly, ${}^{I}x$ (the imaged point on the camera's image plane) is projected to the corresponding virtual image plane as ${}^{I'}x$. This can be done by having the related homography transformation, namely ${}^{I'}H_I$. The corresponding equation is ${}^{I'}x = {}^{I'}H_I\,{}^{I}x$. Eventually, ${}^{I'}x$, the projected point on the virtual image plane, is reprojected to the main reference plane π_{ref} by having the related homography transformation matrix, called ${}^{\pi_{ref}}H_{I'}$. For the first step it is by the camera model described in the Sec. 4. The second and third steps are described in the following two sub-section. Assuming to already have ${}^{I'}H_I$ and ${}^{\pi_{ref}}H_{I'}$, the final equation for registering an image point ${}^{I}x$ onto the reference plane π_{ref} will be (see Fig. 1):

$$ {}^{\pi_{ref}}x = {}^{\pi_{ref}}H_{I'}\;{}^{I'}H_{I}\;{}^{I}x \tag{1}$$

Inertial Compensated Homography: Lets consider ${}^{I''}X$ as re-projection of a point ${}^{I}X$ from image plane *I* of camera *C* onto the virtual image plane *I'* of camera *C'* (see figure 1-right). Then the following general equation can be used:

$$ {}^{I'}x = {}^{I'}H_{I}\;{}^{I}x \tag{2}$$

being ${}^{I'}H_{I}$ a 3×3 homography matrix between real image plane and its virtual plane. In this case H is called *infinite homography* since there is just a pure rotation between real camera and virtual camera centers [6]. In this case the equation for H will become as ${}^{I'}H_{I} = K^{'}RK^{-1}$ [9,13] being R as the rotation matrix between $\{C\}$ and $\{C^{'}\}$ (real and virtual camera centers). Reminding that normal vector of the virtual camera's plane is coincided to the earth gravity vector, then the rotation matrix R can be achieved by using the *IMU* coupled to the camera. The *Camera Inertial Calibration Toolbox* is used in order to calibrate a rigid couple of a IMU and camera [12].

Homography Between Virtual Image Plane and π_{ref}: The homography matrix between the virtual image plane *I'* (the image of virtual camera) and π_{ref} needs to be computed. A homography matrix H can be represented by its axis and vertex and cross-ratio parameters:

$$ H = I + (\mu - 1)\frac{va^{T}}{v^{T}a} \tag{3}$$

in which v is the vertex coordinates and a is the axis line [6,10,13]. For the case of a 3D plane being parallel to the image plane, like in our case, the Eq. 3 becomes a simple matrix by considering $a = (0,0,1)^{T}$ and $v = (v_{x}, v_{y}, 1)$. Then the result is [11]:

$$ {}^{\pi_{ref}}H_{I'} = \begin{bmatrix} 1 & 0 & (\mu-1).v_{x} \\ 0 & 1 & (\mu-1).v_{y} \\ 0 & 0 & \mu \end{bmatrix} \tag{4}$$

which is the homography matrix between the reference plane π_{ref} and virtual image plane I'.

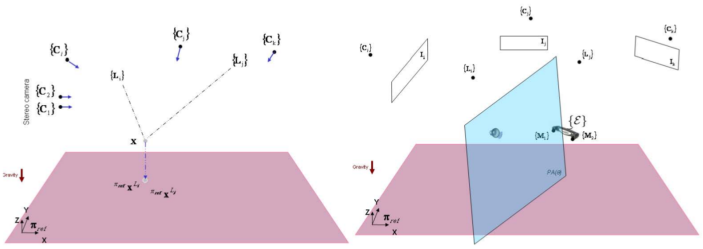

Fig. 2. ***Left***: Range data registration. ***Right***: Registration of azimuth plane of arrival of sound: The intersection of the arrival sound plane and π_{ref} will be registered as a line.

5.2 Range Data Registration

LRF is one of the interesting sensors in our multi-modal setup. The model of our LRF is described in [7]. Geometrical registration of range data is the goal of this section. The general intention is to be able geometrically registering data coming from LRF on the reference plane π_{ref} (see Fig. 2-left). More specifically, for a 3D point X in the scene which is observed in the LRF local reference frame as ${}^{L}x = (x, y, z, 1)$ (in its homogeneous form), its projection on π_{ref}, in homogeneous form ${}^{\pi_{ref}}x = (x, y, 1)$, can be expressed by

$$ {}^{\pi_{ref}}x = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} {}^{W}T_{L}\ {}^{L}x \tag{5} $$

where ${}^{W}T_{L}$ is the transformation matrix between LRF's local reference and can be acomputed by the method described in [7].

5.3 Registration of Azimuth Plane of Arrival of Sound

In this work, we use a simplified version of the sensor model described in Sec. 4, taken the decomposition equation

$$ P(\tau S_{\theta}\theta) = P(\theta)P(S_{\theta} \mid \theta)P(\tau \mid S_{\theta}\theta), \tag{6} $$

where τ is the interaural time-difference binaural cue [9], $\theta \in \{\theta_0, \cdots \theta_n\}$ denotes an azimuth angle taken from a discretised span of n angles, S_θ is a binary variable signaling the existence of a sound-source within the azimuthal plane at θ extending from the frame of reference $\{\xi\}$ (the so-called azimuth plane of arrival for the respective sound-waves), $P(\theta)$ is an uninformative uniform distribution, $P(S_\theta \mid \theta)$ is the prior on the existence of sound-sources for a given θ, also chosen to be a uniform distribution, and $P(\tau \mid S_\theta \theta)$ is the binaural sensor model, a set of normal distributions obtained through calibration (see [8]), which indicate the probability of the measurement of θ (i.e. the angle for the azimuth plane of arrival) knowing whether or not a sound-source is present in that plane. Using Bayes rule it is possible to invert the sensor model so as to obtain $P(S_\theta \mid \tau\theta)$; an auditory saliency map can then be constructed, and a MAP (maximum a posteriori) method can be used to extract the azimuth angle estimate for the most salient sound-source (see Fig. 3-b). Based on these definitions it is possible to also register sound information on the reference plane π_{ref} (see Fig 2). The idea for sound registering is to calculate the intersection between the π_{ref} and the plane of arriving sound. The result will be registered as a line on the π_{ref} .

5.4 Extension for Planes Parallel to π_{ref}

The idea for registering data coming from different sensors onto π_{ref} which was recently described can be extended to register them also on some other planes parallel

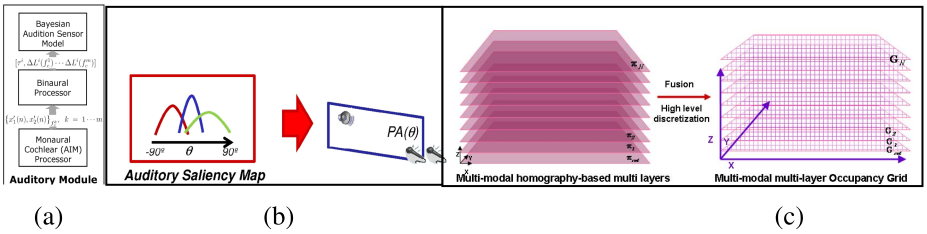
Bayesian Audition Sensor Model
$[\tau^i, \Delta L^i(f_c^1) \cdots \Delta L^i(f_c^m)]$
Binaural Processor
$\{x_1'(n), x_2'(n)\}_{f_c^k}, \ k = 1 \cdots m$
Monaural Cochlear (AIM) Processor
Auditory Module
-90º
θ
90º
Auditory Saliency Map
PA(θ)
Z
Y
X
π_N
π_2
π_1
π_ref
Multi-modal homography-based multi layers
Fusion
High level discretization
G_N
G_2
G_1
G_ref
Multi-modal multi-layer Occupancy Grid
(a)
(b)
(c)

Fig. 3. (a): Bayesian binaural system. **(b)**: Auditory saliency map and azimuth plane of arrival extracted for the most salient sound source. **(c)**: Multi-modal multi-layer data registration and data fusion framework.

to π_{ref}. Fig. *3-c-left* represents a setup with *N-1* planes parallel to π_{ref}. Such an extension needs to be done for cameras, LRFs and microphones. For LRF it becomes simple since a general method to find the homogeneous transformation between LRF and $\{W\}$ has been defined in Sec. 5.2. For arrival sound it is exactly the same method described in Sec. 5.3. Here we continue to describe this extension for images. The intention is to calculate homography for any of these layers having the homography of the reference layer $^{\pi_{ref}}H_{I}$. For doing so we use *Khan*'s approach described in [5].

5.5 Multi-layer Homography-Based Occupancy Grid

Fig. 3-c shows a schematic of the proposed multi-layer framework for data registration. The idea is to make an occupancy grid layer for each plane of the registration framework. In order to make the final data fusion a BOF grid [14] per each plane is proposed as an occupancy layer (see Fig. 3-(c)). It can be used as input of a classification algorithm in order to classify the objects inside the scene and then be used by a tracking algorithm. We use the concept and also the model described in [14] for using BOF in our case. Here the variables definition is the same as [14]. Using that the probability joint distribution for the model becomes also the same as what is defined in [14] (*l* being the layer index):

$$P({}^{l}A_{c^l}^{t-1}\ {}^{l}A_{c^l}^{t}\ O_{c^l}^{t}\ Z_1^t \ldots Z_S^t) = P({}^{l}A_{c^l}^{t-1})P(P({}^{l}A_{c^l}^{t} \mid {}^{l}A_{c^l}^{t-1})P(O_{c^l}^{t} \mid {}^{l}A_{c^l}^{t-1})\prod_{i=1}^{S}P(Z_i^t \mid {}^{l}A_{c^l}^{t}\, O) \tag{7}$$

6 Experiments

As a preliminary experiment result the image data registration part is performed for just one layer. We used the CVLab - EPFL dataset [15]. Fig. 4:(a)-(d) show four images from four views and their projection on the ground plane using homography concept. Then after background subtraction, they are combined and intersected in order to obtain the feet positions on the ground plane (Fig. 4-e).

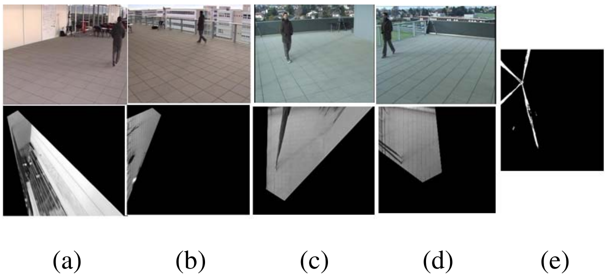

Fig. 4. (a) to (d): Images from different views and their projection on the ground plane using homography concept. e: combining four ground-projected images in order to have the feet intersection.

7 Conclusion and Future Work

Data registration and data fusion have been investigated in this paper where there are four heterogeneous sensors such as camera, microphone, laser range finder and inertial sensor. Then a multi-modal multi-layer framework is porposed. In such a framework firstly sensory data coming from different sensors are geometrically registered on different parallel planes, then an occupancy grid is applied for each layer. A prelimanry experiment result has been shown for image data registration in just one layer in Sec. 6. Data registration for other layers, range and sound data and also constructing an occupancy grid for each layer will be remained as our future work.

Acknowledgment. *Hadi Ali Akbarpour* and *Joao Filipe Ferreira* are supported by the **FCT** (Portuguese Fundation for Sceince and Technology). This work is partially supported by the European Union within the FP7 Project **PROMETHEUS**, www.prometheus-FP7.eu.

References

1. Smith, D., Singh, S.: Approaches to multisensor data fusion in target tracking: A survey. IEEE Transactions on Knowledge and Data Engineering 18, 1696–1710 (2006)
2. Armesto, L., Tornero, J.: On multi-rate fusion for non-linear sampled-data systems: Application to a 6d tracking system. Robotics and Autonomous Systems. Elsevier, Amsterdam (2007)
3. Bellotto, N., Hu, H.: Vision and laser data fusion for tracking people with a mobile robot. In: Proceedings of the 2006 IEEE International Conference on Robotics and Biomimetics, Kunming, China (2006)
4. Chakravarty, J.: Panoramic vision and laser range finder fusion for multiple person tracking. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS (2006)
5. Khan, S.M., Yan, P., Shah, M.: A homographic framework for the fusion of multi-view silhouettes. In: IEEE 11th International Conference on Computer Vision, ICCV 2007 (2007)

6. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2003)
7. Aliakbarpour, H., Nunez, P., Prado, J., Khoshhal, K., Dias, J.: An efficient algorithm for extrinsic calibration between a 3d laser range finder and a stereo camera for surveillance. In: 14th International Conference on Advanced Robotics, ICAR 2009 (2009)
8. Ferreira, J.F., Pinho, C., Dias, J.: Implementation and calibration of a Bayesian binaural system for 3d localisation. In: IEEE International Conference on Robotics and Biomimetics (ROBIO 2008), Bangkok, Tailand, December 2008, pp. 14–17 (2008)
9. Pinho, C., Ferreira, J.F., BessiÃšre, P., Dias, J.: A Bayesian binaural system for 3d sound-source localisation. In: International Conference on Cognitive Systems (CogSys 2008), pp. 109–114 (2008)
10. Mirisola, L.G.B., Dias, J.: Tracking from a moving camera with attitude estimates. In: ICR 2008 (2008)
11. Mirisola, L.G.B.: Exploiting attitude sensing in vision-based navigation, mapping and tracking including results from an airship. PhD thesis (2009)
12. Lobo, J., Dias, J.: Relative pose calibration between visual and inertial sensors. International Journal of Robotics Research, Special Issue 2nd Workshop on Integration of Vision and Inertial Sensors 26, 561–575 (2007)
13. Criminisi, A.: Accurate visual metrology from single and multiple uncalibrated images. PhD thesis, Oxford (1999)
14. Mekhnacha, K., Mao, Y., Raulo, D., Laugier, C.: Bayesian occupancy filter based fast clustering-tracking algorithm. In: IROS 2008 (2008)
15. Fleuret, F., Jerome Berclaz, R.L., Fua, P.: Multi-camera people tracking with a probabilistic occupancy map. IEEE Transactions on Pattern Analysis and Machine Intelligence (2008)