Multiple camera fusion based on DSmT for tracking objects on ground plane

Esteban Garcia and Leopoldo Altamirano
National Institute for Astrophysics, Optics and Electronics
Puebla, Mexico
eomargr@inaoep.mx, robles@inaoep.mx

***Abstract*—This paper presents comparative results of a model for multiple camera fusion, which is based on Dezert-Smarandache theory of evidence. Our architecture works at the decision level to track objects on a ground plane using predefined zones, producing useful information for surveillance tasks such as behavior recognition. Decisions from cameras are generated by applying a perspective-based basic belief assignment function, which represent uncertainty derived from cameras perspective while tracking objects on ground plane. Results obtained from applying our tracking model to CGI animated simulations and real sequences are compared to the ones obtained by Bayesian fusion, and show how DSm theory of evidence overcomes Bayesian fusion for this application.**

Keywords: Multiple Cameras Fusion, Tracking, Dezert-Smarandache Theory, Decision Level Fusion.

I. Introduction

Computer vision uses information from more than one camera to develop several tasks, such as 3D reconstruction or complementing fields of view to increase surveillance areas, among others. Using more than one camera has some advantages, even if information is not fused. A simple instance might be having a multi-camera system where it is possible to cover wider area, and at the same time is more robust to failures where cameras overlap.

There exists a tendency, in computer vision, to work on high level tasks [1]–[4], where moving objects position is not useful when it is given in image plane coordinates, instead of it, it is prefered when position is described according to predefined regions on ground plane. This sort of information can be used for behavior recognition where people behavior is described by mean of predefined zones of interest on scene.

In [4] a tracking system using predefined regions is used to analyze behavioral patterns. In the same work, only one camera is used and no considerations are taken on distortions due to camera perspective. In [3] a Hierarchical Hidden Markov Model is used to identify activities, based on tracking people on a cell divided room. Two static cameras cover scene, but information coming from them is used separately, their purpose is to focus on different zones, but not to refine information.

As cameras work by transforming information from 3D space into 2D space, there is always uncertainty involved. In order to estimate object position related to ground plane, it is necessary to find out its position in image plane and then estimate that position on ground plane. For surveillance tasks where objects position has to be given according to ground plane, it is possible to apply projective transform in order to estimate objects position on ground plane, however, this process might carry errors from perspective.

In [5] we presented a decision level architecture to fuse information from cameras, reducing uncertainty derived from perspective on cameras. The stage of the processing at which data integration takes place allows an interpretation of information which describes better the position of objects being observed and at the same time is useful for high level surveillance systems. In our proposal, individual decisions are taken by means of an axis-projection-based *generalized basic belief assignment* (gbba) function and finally fused using Dezert-Smarandache (DSm) hybrid rule. In this work, we present a theoretical and practical comparison between DSm and a Bayesian module applied to CGI and real multicamera sequences.

This paper is organized as follows: in section 2, the Dezert-Smarandache theory is briefly described as mathematical framework. In section 3, our architecture is described altogether with the gbba function we used. A comparison between Bayesian and DSm hybrid combination rule is presented in section 4. Finally in section 5 conclusions are presented.

II. DSm hybrid model

The DSmT defines two mathematical models used to represent and combine information [6]: free and hybrid.

The *Free DSm model*, denoted as $\mathcal{M}^f(\Theta)$, defines $\Theta = \{\theta_1, \ldots, \theta_n\}$ as a set or frame of n non exclusive elements and an hyper-power set D^Θ as the set of all composite possibilities obtained from Θ in the following way:

1) $\emptyset, \theta_1, \ldots, \theta_n \in D^\Theta$
2) $\forall A \in D^\Theta, B \in D^\Theta, (A \cup B) \in D^\Theta, (A \cap B) \in D^\Theta$
3) D^Θ is formed only by elements obtained by rules 1 or 2

Function $m(A)$ is called general basic belief assignment or mass for A, defined as $m() : D^\Theta \rightarrow [0, 1]$, and is associated to a source of evidence.

A DSm hybrid model introduces some integrity constraints on elements $A \in D^\Theta$ when there are known facts related to those elements in the problem under consideration. In our work, exclusivity constraints are used to represent those regions on ground plane which are not adjacent. The restricted elements are forced to be empty in the hybrid model $\mathcal{M}(\Theta) \neq \mathcal{M}^f(\Theta)$

Figure 1. Example of vertical axis obtained by two cameras, projected on ground plane

and the mass is transferred to the non restricted elements. When DSm hybrid model is used, combination rule for two or more sources is defined for $A \in D^\Theta$ with these functions:

$$m_{\mathcal{M}(\Theta)}(A) = \phi(A)\left[S_1(A) + S_2(A) + S_3(A)\right] \quad (1)$$

$$S_1(A) = \sum_{\substack{X_1,X_2,\ldots,X_k \in D^\Theta \\ X_1 \cap X_2 \cap \ldots \cap X_k = A}} \prod_{i=1}^{k} m_i(X_i) \quad (2)$$

$$S_2(A) = \sum_{\substack{X_1,X_2,\ldots,X_k \in \emptyset \\ [\mathcal{U}=A] \vee [[\mathcal{U} \in \emptyset] \wedge [A = I_t]]}} \prod_{i=1}^{k} m_i(X_i) \quad (3)$$

$$S_3(A) = \sum_{\substack{X_1,X_2,\ldots,X_k \in D^\Theta \\ X_1 \cup X_2 \cup \ldots \cup X_k = A \\ X_1 \cap X_2 \cap \ldots \cap X_k \in \emptyset}} \prod_{i=1}^{k} m_i(X_i) \quad (4)$$

where $\phi(A)$ is called the characteristic emptiness function of a set A ($\phi(A) = 1$ if $A \notin \mathbf{\emptyset}$ and $\phi(A) = 0$ otherwise). $\mathbf{\emptyset} = \{\emptyset_\mathcal{M}, \emptyset\}$ where $\emptyset_\mathcal{M}$ is the set of all elements of D^Θ forced to be empty. $\mathcal{U}$ is defined as $\mathcal{U} = u(X_1) \cup u(X_2) \cup \ldots \cup u(X_k)$, where $u(X)$ is the union of all singletons $\theta_i \in X$, while $I_t = \theta_1 \cup \theta_2 \cup \ldots \cup \theta_n$.

III. Multiple cameras fusion

In order to have a common space reference system, spatial alignment is required. Homography is used to relate information from cameras. It is possible to recover homography from a set of static points on ground plane [7] or dynamic information in scene [8]. Correspondence between objects detected in cameras might be achieved by features matching techniques [9] or geometric ones [10], [11].

Once the homography matrix has been calculated, it is possible to relate information from one camera to others. While object is being tracked by a camera, its vertical axis is obtained and its length is estimated as $\lambda = l\cos(\alpha)$, where l is the maximum length for axis when projected on ground plane and α is the angle of the camera respect to the ground plane.

Let $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$ denote ground plane partition, where each γ_x is a predefined region on ground plane, which might be an special interest zone, such as corridor or parking area.

For each moving object i, it is created a frame $\Theta_i = \{\theta_1, \ldots, \theta_k\}$. Each element θ_x represents a zone γ_y where the object i might be located, according to information from cameras. Θ_i is built dynamically considering only the zones for which there exist some belief provided by at least one camera.

Multiple camera fusion, in the way it is used in this work, is a tool for high level surveillance systems. Behavior recognition models might use information in the form of beliefs, such as fuzzy logic classifiers or probabilistic models do. Therefore, it is allowed for the camera to assign mass to elements in D^Θ in the form of $\theta_i \cap \theta_j$, because this might represent an object in the border of two regions on ground plane. For couples of hypotheses which represent non-adjacent regions of the ground plane, it does not make sense consider such belief assignments, therefore elements in D^Θ representing non-adjacent regions of ground plane, are included to $\mathbf{\emptyset}_\mathcal{M}$.

Each camera behaves as an expert, assigning mass to each one of the unconstrained elements of D^Θ. The assignment function is simple, and has as its main purpose to consider perspective influence on uncertainty. It is achieved by means of measuring intersection area between γ_x and object's vertical axis projected on ground plane, centered on the object's feet. The length of the axis projected on ground plane is determined by the angle of the camera respect to the ground plane, taking object's ground point as the vertex to measure the angle. So if the camera were just above the object, its axis projection would be just one pixel long, meaning no uncertainty at all. We consider three cases to cover mass assignation showed in figure 2.

When projected axis is within a region of the ground plane, camera assigns full belief to that hypothesis. When axis crosses two regions it is possible to assign to composed hypotheses of the kind $\theta_i \cup \theta_j$ and $\theta_i \cap \theta_j$, depending on the angle of the camera.

Let ω_c denote the vertical axis obtained by camera c, projected on ground plane, and $|\omega_c|$ its area. Following functions are used as gbba model.

$$\upsilon = |\omega_c|\cos(\alpha_c) \quad (5)$$

$$m_c(\theta_i) = \frac{|\omega_c \cap \gamma_x|}{\upsilon + |\omega_c|} \quad (6)$$

$$m_c(\theta_i \cup \theta_j) = \frac{|\omega_c|\cos^2(\alpha_c)}{\upsilon + |\omega_c|} \quad (7)$$

$$m_c(\theta_i \cap \theta_j) = \frac{\upsilon(1 - \cos(\alpha_c))}{\upsilon + |\omega_c|} \quad (8)$$

When axis intersects more than two regions on ground plane, functions become:

(a) Belief is assigned to θ_i

(b) Belief is assigned to θ_i, θ_j, $\theta_i \cup \theta_j$ and $\theta_i \cap \theta_j$

(c) Belief is assigned to θ_i, ... , θ_k and $\theta_i \cup \ldots \cup \theta_k$

Figure 2. Cases considered to belief assignment

$$\upsilon = |\omega_c|\cos(\alpha_c) \quad (9)$$

$$m_c(\theta_i) = \frac{|\omega_c \cap \gamma_x|}{\upsilon + |\omega_c|} \quad (10)$$

$$m_c(\theta_i \cup \theta_j \cup \ldots \cup \theta_k) = \frac{\upsilon}{\upsilon + |\omega_c|} \quad (11)$$

$\upsilon + |\omega_c|$ is used as a normalizer in order to satisfy $m_c() \rightarrow [0,1]$ and Each camera can provide belief to elements $\theta_x \cap \theta_y \in D^\Theta$, by considering couples γ_i and γ_j (represented by θ_x and θ_y respectively) crossed by axis projection. Elements $\theta_x \cup \ldots \cup \theta_x$ can have an associated gbba value, which represents local or global ignorance. We also restrict elements in $\theta_x \cap \ldots \cap \theta_y \in D^\Theta$ for which there is not a direct basic assignation made by one of the cameras, thus they are included in $\emptyset_\mathcal{M}$, and calculations are simplified. That is possible because of the hybrid DSm model definition.
Decision fusion is used to combine the outcomes from cameras, making a final decision. We apply hybrid DSm rule of combination over D^Θ in order to achieve a final decision.

IV. Results and discussion

To test the proposed architecture for fusion, we used computer-generated-imagery sequences (figure 4) and real sequences from the Performance Evaluation of Tracking and Surveillance dataset [12].

In CGI sequences, three cameras were simulated. We considered a squared scenario with a grid of sixteen regular predefined zones. 3D modeling was done using Blender with Yafray as rendering machine. All generated images for sequence are in a resolution of 800x600 pixels. Examples of images generated by rendering are shown in figure 4, where division lines were outlined on ground plane to have a visual reference of zones, but they are not required for any other task.

As real sequences, PETS repository was used (figure 5). In this data set, two cameras information is provided, in a resolution of 768x576 pixels in JPEG format. Our architecture and gbba function was applied to track people, cars and bicycles.

As part of the results, it is interesting to show the differences between DSm and a probabilistic model to fuse decisions. For this application, hypotheses have a geometric meaning, and we found that this has to be taken in consideration during fusion.

A. Probabilistic fusion module

Figure 3. Bayesian classifiers as fusion module

For comparison purposes, a Bayesian classifier was developed for each of the regions on ground plane, as showed in figure 3. *A priori* probability is assumed the same for each of the regions, while conditional probability is taken from masses generated by cameras, being normalized.

$$p(\gamma_i|S_1,\ldots,S_n) = \frac{p(\gamma_i)p(S_1,\ldots,S_n|\gamma_i)}{p(S_1,\ldots,S_n)}$$

$$\begin{aligned} p(\gamma_i|S_1,\ldots,S_n) &\propto p(\gamma_i)p(S_1|\gamma_i)p(S_2|\gamma_i)p(S_3|\gamma_i)\ldots \\ &= p(\gamma_i)\prod_{i=1}^{n} p(S_i|\gamma_i) \end{aligned}$$

Ignorance from cameras means that a camera does not have a good point of view to generate its information. If a probabilistic model is applied ignorance is not considered and that might derive wrong results. Let's consider the following numerical example: suppose two cameras assign following beliefs:

$$m_1(A) = 0.35\ m_1(B) = 0.6\ m_1(A \cup B) = 0.05$$
$$m_2(A) = 0.3\ m_2(B) = 0.1\ m_2(A \cup B) = 0.6$$

Probabilistic model generates following decisions:

(a) Camera 1 (b) Camera 2 (c) Camera 3

Figure 4. Example of CGI sequences

(a) Camera 1 (b) Camera 2 (c) Ground plane

Figure 5. Example of real sequences from PETS

$$p(A) \propto 0.5 \cdot \frac{0.35}{0.35+0.6} \cdot \frac{0.3}{0.3+0.1} = 0.13$$
$$p(B) \propto 0.5 \cdot \frac{0.6}{0.35+0.6} \cdot \frac{0.1}{0.3+0.1} = 0.07$$

DSm model results:

$$m_{DSm}(A) = 0.35 \cdot 0.3 + 0.35 \cdot 0.6 + 0.05 \cdot 0.3 = 0.33$$
$$m_{DSm}(B) = 0.6 \cdot 0.1 + 0.6 \cdot 0.6 + 0.05 \cdot 0.1 = 0.42$$

In decisions generated by cameras, first sensor assign higher mass to the hypothesis B, while second sensor assigns higher belief to hypothesis A. If ignorance is considered, it is clear that as result from fusion one must have a higher value for hypothesis B, because second sensor is in a better position. However, in probabilistic fusion decision hypothesis A is higher. This shows how considering ignorance may improve results from fusion applied to multi-cameras tracking.

Positions obtained by fusion of the decisions of the cameras are showed in figures 6 and 7. Graphics show how DSm gets higher decision values than Bayesian fusion.

In tables I and II metrics TRDR (Tracker Detection Rate) and FAR (False Alarm Rate) are showed from data collected from 2 CGI sequences and 5 real sequences. We also propose *Similarity to Truth* measure, to evaluate how close in values is the result of fusion to truth data.

$TRDR$ and FAR are evaluated with following equations:

$$\text{TRDR} = \frac{TP}{TG} \tag{12}$$

$$\text{FAR} = \frac{FP}{TP+FP} \tag{13}$$

where TG is the total number of regions by each image where there are objects in motion according to ground truth. According to this metrics, it is desirable to have the highest value in $TRDR$ while the lowest in FAR.

Similarity to Truth is a measure to quantify the differences between positions obtained by fusion modules compared to ground truth. When there exist belief assigned to certain position, and also exists an object on that position in ground truth, the amount of belief is summed, but when there is not object in ground truth, this amount of belief is subtracted, and finally, the amount obtained is normalized to be showed as percentage.

Table I
RESULTS ON CGI ANIMATIONS

Source	TRDR	FAR	Similarity to Truth
Camera 1	99.5%	52.9%	65.2%
Camera 2	93.9%	43.0%	69.7%
Camera 3	84.4%	45.3%	23.0%
DSm	93.9%	5.6%	84.1%
Probabilistic	93.3%	5.2%	24.9%

Table II
RESULTS ON REAL SEQUENCES

Source	TRDR	FAR	Similarity to Truth
Camera 1	68.1%	21.7%	31.6%
Camera 2	71.0%	2.7%	67.5%
DSm	82.8%	10.2%	75.9%
Probabilistic	82.8%	10.2%	67.9%

Results from tables show how DSm reduces uncertainty from perspective and complements information where cameras lost object or fields of view do not overlap. Bayesian fusion behaves similar to DSm, however, hybrid combination rule takes in consideration information assigned to ignorance, which may refine information such as in example from section IV-A. ST (Similarity to Truth) is a metric to quantify how close is belief assigned to regions to ground truth. In ST DSm has higher values, closer to ground truth.

V. CONCLUSIONS

Using cameras as experts at high level for processing objects position, allows to apply Dezert-Smarandache Theory to combine beliefs. Beliefs correspond to objects locations on ground plane, given in relation to predefined regions.

Test showed how DSm Theory of evidence generates higher values as results and a better approximation to ground truth. In addition to this, DSmT allows belief to be assigned to intersection of hypotheses, which might be interpreted as an object in the border of two regions, and might be useful information for behavior recognition based on fuzzy logic, while probabilistic approaches does not allow this kind of information because of exclusivity constraints. For the fusion of objects position, DSmT showed better results than Bayesian fusion.

Even good results were obtained using DSmH, it is known that when conflicting sources are combined the masses committed to partial ignorances are increased and after a while this ends up to get the vacuous belief assignment. It is expected that DSm-PCR5 fusion rule yelds better results.

REFERENCES

[1] F. Lv, J. Kang, R. Nevatia, I. Cohen, and G. Medioni, "Automatic tracking and labeling of human activities in a video sequence," IEEE International Workshop on Performance Evaluation of Tracking and

(a) True position

(b) Decisions by DSm

(c) Decisions by Bayesian fusion

Figure 6. Example of positions obtained in 3D animations. Belief value is plotted from blue to red, blue meaning low belief and red meaning 1.

(a) True position

(b) Decisions by DSm

(c) Decisions by Bayesian fusion

Figure 7. Example of positions obtained in real sequences. Belief value is plotted from blue to red, blue meaning low belief and red meaning 1.

Surveillance In conjunction with ECCV'04, 2004. [Online]. Available: citeseer.ist.psu.edu/lv04automatic.html

[2] R. Green and L. Guan, "Quantifying and recognizing human movement patterns from monocular video images - part i: A new framework for modeling human motion," 2003. [Online]. Available: citeseer.ist.psu.edu/article/green03quantifying.html

[3] N. T. Nguyen, D. Q. Phung, S. Venkatesh, and H. Bui, "Learning and detecting activities from movement trajectories using the hierarchical hidden markov models," in *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 955–960.

[4] W. Yan and D. A. Forsyth, "Learning the behavior of users in a public space through video tracking," in *WACV-MOTION '05: Proceedings of the Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION'05) - Volume 1*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 370–377.

[5] E. Garcia and L. A. Robles, "Decision level multiple cameras fusion using dezert-smarandache theory," in *CAIP*, ser. Lecture Notes in Computer Science, W. G. Kropatsch, M. Kampel, and A. Hanbury, Eds., vol. 4673. Springer, 2007, pp. 117–124.

[6] F. Smarandache and J. Dezert, "An introduction to the dsm theory for the combination of paradoxical, uncertain, and imprecise sources of information," 2006. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:cs/0608002

[7] G. Stein, L. Lee, and R. Romano, "Monitoring activities from multiple video streams: Establishing a common coordinate frame," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 758–767, 2000.

[8] K. J. Bradshaw, I. D. Reid, and D. W. Murray, "The active recovery of 3d motion trajectories and their use in prediction." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 3, pp. 219–234, 1997.

[9] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multiperson tracking for easyliving," in *Proceedings of the Third IEEE International Workshop on Visual Surveillance*, July 2000, pp. 3–10.

[10] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. vol. 25, no. 10, pp. pages 1355–1360, 2003.

[11] J. Black and T. Ellis, "Multi camera image tracking." *Image Vision Comput.*, vol. 24, no. 11, pp. 1256–1267, 2006.

[12] http://www.cvg.cs.rdg.ac.uk/PETS2001/pets2001 dataset.html, "Performance evaluation of tracking and surveillance dataset," Internet Reference.