Dempster-Shafer based multi-view occupancy maps

Marleen Morbee[1], Linda Tessens[1], Hamid Aghajan[2], Wilfried Philips[1]

[1]Ghent University, TELIN-IPI-IBBT, Sint-Pietersnieuwstraat 41, Ghent, Belgium

[2]Stanford University, Department of Electrical Engineering, WSNL, Stanford, CA, USA

{marleen.morbee,linda.tessens}@telin.UGent.be

Abstract

We present a novel method for calculating occupancy maps with a set of calibrated and synchronised cameras. In particular, we propose Dempster-Shafer based fusion of the ground occupancies computed from each view. The method yields very accurate occupancy detection results and in terms of concentration of the occupancy evidence around ground truth person positions it outperforms the state-of-the-art probabilistic occupancy map method and fusion by summing.

1 Introduction

An occupancy map provides a top view of a scene containing people or objects. Such maps are important in many applications such as surveillance, smart rooms, video conferencing and sport games analysis. Camera networks offer an attractive non-intrusive and flexible tool for this purpose. They do not require people to wear dedicated gear, nor the environment to

be equipped with special sensors other than cameras, which are often part of the existing infrastructure, especially in security applications.

In recent years, foreground silhouettes in multiple camera views have been increasingly used to estimate the probability of ground occupancy. Two basic approaches exist. Bottom-up methods transfer the information in the different camera images to a common reference plane using camera image-floor homographies [1]. Top-down approaches extract occupied ground positions by comparing a generative model of the objects in the scene with the actual foreground silhouettes observed in the camera views [2, 3]. Until now, for both approaches the mathematical laws for the fusion of data from different cameras have not been considered explicitly. In this letter we focus on this data fusion aspect within a bottom-up method and show that Dempster-Shafer based fusion of camera information leads to significantly more accurate occupancy maps. For the basket ball dataset of [4], the total mass of occupancy evidence is 1.08 to 8.13 times more concentrated around the ground truth player positions than for the methods of [1] and [2], as will be discussed in Section 4.

2 Data fusion

In the probabilistic occupancy map (POM) method of [2], for each view the conditional distribution of the observed background subtraction image given the true object positions is a function of a distance measure between the background subtraction image and the image obtained from a generative model. Information from different views is fused by multiplying these conditional distributions. This strategy is problematic in the typical case of imperfect foreground detection: a missed foreground region in even a single view can easily result in a missed occupancy detection.

In [1], each camera produces a confidence value for the occupancy of each ground position

by back-projecting the foreground silhouettes to a common reference plane using camera image-floor homographies. The aggregated ground occupancy map is obtained by summing the camera confidences and by normalizing by the number of cameras that actually view $\mathbf{x}$.

In this letter, unlike the summing [1] and POM [2] fusion strategy, we use Dempster-Shafer (DS) based fusion to exploit the fact that if a hypothesis of (non-)occupancy is corroborated by different cameras, a higher belief should be assigned to it. Moreover, the DS theory of evidence allows distinguishing between equal probability of occupancy and non-occupancy, and lack of knowledge, e.g. when an object is outside a camera's viewing range. More specifically, in our method the cameras are considered independent sources of information whose data about the (non-)occupancy of ground positions can be opportunistically fused using the DS rule of combination [5].

3 Occupancy maps

3.1 Problem Formulation

Consider a network of N cameras and let the ground plane of the observed scene be discretised in resolution cells $\mathbf{x}$. The discretisation resolution should be chosen such that the area covered by one cell is (typically a lot) smaller than the average area occupied by a person. We wish to assign a real value to each cell that expresses our confidence that the cell is occupied.

3.2 Proposed Method

In the DS theory of evidence, a basic belief assignment or BBA m is a mapping that assigns to each subset A of a frame of discernment $\theta_{\mathbf{x}}$ a belief $m(A) \in [0, 1]$. The total assigned belief should be 1 and the belief of the empty set $\emptyset$ should be 0. The basic belief assigned

to a hypothesis expresses how much evidence supports it. In our method, for each cell $\mathbf{x}$ the mutually exclusive and exhaustive hypotheses that $\mathbf{x}$ is either occupied ($\{occ_{\mathbf{x}}\}$) or not ($\{nocc_{\mathbf{x}}\}$) constitute the frame of discernment $\theta_{\mathbf{x}} = \{occ_{\mathbf{x}}, nocc_{\mathbf{x}}\}$ [5]. The information from each view i, $1 \leq i \leq N$, is considered a distinct piece of evidence and we denote the BBA representing this evidence by m_i. We now explain how we define the BBA in our method.

Let H be the typical height of a person and consider a rectangular cuboid with cell $\mathbf{x}$ as base and height H. If this cuboid lies completely outside the viewing frustum of camera i, this camera cannot provide any information about the occupancy of $\mathbf{x}$. The BBA is then $m_i(\{occ_{\mathbf{x}}\}) = 0$, $m_i(\{nocc_{\mathbf{x}}\}) = 0$ and $m_i(\theta_{\mathbf{x}}) = 1$. Otherwise, the projection of this cuboid into camera view i defines an image region $R^i_{\mathbf{x}}$. We gather evidence about the (non-)occupancy of the cells by independently segmenting each view into background and foreground, and by determining in each region $R^i_{\mathbf{x}}$ the fraction of background pixels $b^i_{\mathbf{x}}$ and of foreground pixels $f^i_{\mathbf{x}}$. Of course $b^i_{\mathbf{x}} + f^i_{\mathbf{x}} = 1$. The evidence $m_i(\{nocc_{\mathbf{x}}\})$ of camera i for the hypothesis $\{nocc_{\mathbf{x}}\}$ is $b^i_{\mathbf{x}}$.

For $m_i(\{occ_{\mathbf{x}}\})$ the situation is more complicated: because of the limited resolution of the cameras, different cells $\mathbf{x}$ and $\mathbf{x}'$ may give rise to completely coinciding regions $R^i_{\mathbf{x}}$ and $R^i_{\mathbf{x}'}$. Let $G^i_{\mathbf{x}}$ be the number of cells sharing the same region $R^i_{\mathbf{x}}$. If $G^i_{\mathbf{x}} > 1$, the evidence of occupancy collected in $R^i_{\mathbf{x}}$ may be attributable to a person occupying only part of the cells with coinciding $R^i_{\mathbf{x}}$. Because of the reprojection geometry, these $G^i_{\mathbf{x}}$ positions will be approximately laid out in a trapezoid, which we approximate by a square $\mathbf{S}$ with side length $\sqrt{G^i_{\mathbf{x}}}$.

Assuming a person occupies a square of W^2 cells, this person can be in $(\sqrt{G^i_{\mathbf{x}}} + W - 1)^2$ different positions with respect to the square $\mathbf{S}$. A particular cell $\mathbf{x}$ in the square $\mathbf{S}$ is only occupied in W^2 of all these positions. Hence, the evidence of occupancy $m_i(\{occ_{\mathbf{x}}\})$ is scaled with $g^i_{\mathbf{x}} = W^2/(\sqrt{G^i_{\mathbf{x}}} + W - 1)^2$ and $m_i(\{occ_{\mathbf{x}}\}) = g^i_{\mathbf{x}} f^i_{\mathbf{x}}$.

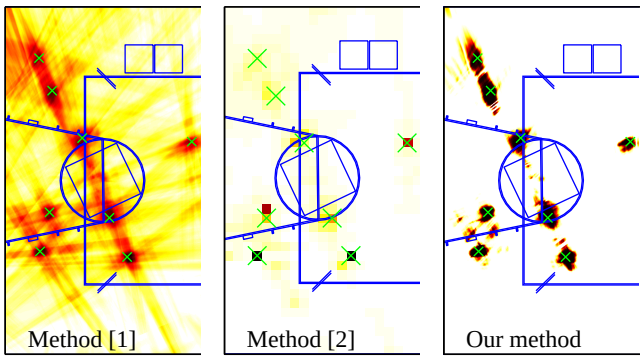

Figure 1: Left the aggregated [1], in the middle the probabilistic [2] and right the proposed evidential occupancy map. White corresponds to low confidence/probability/evidence of occupancy, black to high. The crosses indicate the ground truth player positions.

With $m_i(\{occ_{\mathbf{x}}\})$ and $m_i(\{nocc_{\mathbf{x}}\})$ defined, $m_i(\theta_{\mathbf{x}}) = 1 - m_i(\{occ_{\mathbf{x}}\}) - m_i(\{nocc_{\mathbf{x}}\})$.

The evidences collected by the N views about each cell $\mathbf{x}$ are fused using Dempster's rule of combination [5]. The fused evidence of occupancy for all cells forms an occupancy map which we denote as $m(\{occ\})$.

4 Results

To evaluate our method, we use the publicly available basketball dataset from the European project APIDIS [4]. It consists of seven synchronised and calibrated video streams from five cameras with partially overlapping views distributed around the court, and two top-mounted cameras with fish eye lenses. The videos are captured at 2 megapixel resolution and 25 fps. The size of the field is $15m \times 28m$. There are on average 12 targets on the field. The average height of a player is set to $2m$, as in [1]. We consider square resolution cells with an area of $(0.02m)^2$. In the rare case of conflicts in the fusion process, all evidence is transferred to $m(\theta_{\mathbf{x}})$. The foreground is detected using an algorithm based on mixture of Gaussians modelling with elementary shadow removal.

Ground truth target positions have been made available for 60 frames recorded at 1 sec intervals within the time interval 18:47 until 18:48 [4]. As most cameras point to the left half of the court, only positions in that half are considered for the evaluation.

The right panel of Fig. 1 shows an example of $m(\{occ\})$ in part of the left half of the court. The left panel in Fig. 1 shows the aggregated occupancy map obtained as in [1], the middle one the probabilistic occupancy map of [2] with cell width set to $0.4m$ (other widths yield less accurate results). The map obtained by DS fusion is more representative of the actual occupancy of the field because it shows very clearly defined peaks at the target positions, and very few ghost objects or interference strokes between objects. This is less the case for the methods of [1] and [2].

Let the total mass (TM) be the sum over all cells of the occupancy evidence for the proposed method ($\mathrm{TM} = \sum_{\forall \mathbf{x}} \mathrm{m}(\{\mathrm{occ}_{\mathbf{x}}\})$), of the aggregated occupancy confidence for the method of [1], and of the occupancy probability for the method of [2]. In Fig. 2, we plot for our method and the method of [1] the percentage of TM that lies within a disc with diameter d around a ground truth target position as a function of d. For the method of [2] this evaluation method yields poor results since in this method the correlation between the occupancy probability of adjacent cells is explicitly ignored. Hence, to obtain good results with this method the size of the resolution cells should approximate the expected size of the objects to detect and this cell size is significantly larger than in our method and the method of [1]. Therefore, for fair comparison we plot for the method of [2] for different cell widths d the percentage of TM that is generated in cells that are actually occupied by a target.

From this graph we conclude that in the proposed method the mass of occupancy evidence is more concentrated around the ground truth positions than the mass of occupancy confidence of method [1] and the mass of occupancy probability of method [2]. This is obvious from the ratio between the percentage of total mass of our method and the method of [1] and [2].

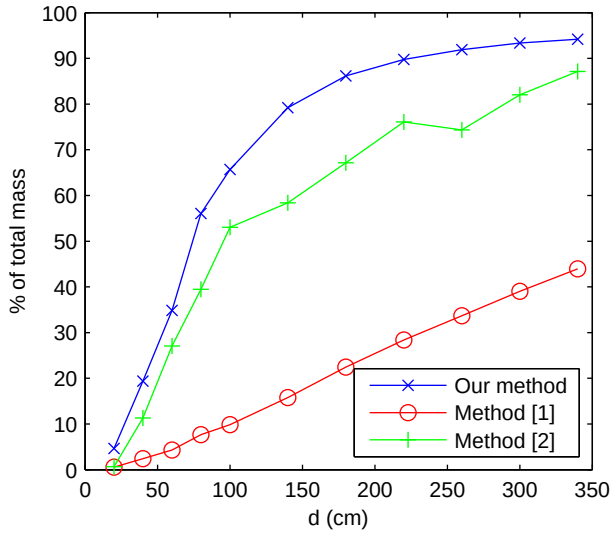

Figure 2: The percentage of the total mass within a disc with diameter d around a ground truth target position (for the proposed method and the method of [1]), or within cells with width d actually occupied by a target (for method [2]).

For [1], this ratio ranges from $19.38\%/2.38\% = 8.13$ for $d = 40$cm to $94.18\%/43.96\% = 2.14$ for $d = 340$cm, and reaches 6.66 for a typical diameter of $1m$ for sports players. For [2], it ranges from $4.65\%/0.65\% = 7.13$ for $d = 20$cm to $94.18\%/87.15\% = 1.08$ for $d = 340$cm, and reaches 1.24 for $d = 1m$. In other words, the ground occupancy map obtained using the proposed method is more accurate than using the methods of [1] and [2]. This is beneficial for direct use or for further analysis of the map.

The proposed method is about a factor of 6 more complex than the method of [1]. Indeed, fusing the bodies of evidence of two cameras requires 17 operations per cell. For N cameras this boils down to $17(N-1)$ operations, compared to $3N+1$ operations required for [1]. Due to the iterative nature of the algorithm of [2], its complexity is a factor in the order of hundreds higher than that of the proposed method.

5 Conclusion

We have described a new method to calculate occupancy maps using multiple cameras. In particular, we have shown how the performance of a method requiring only forward projections from the image to the ground plane can be significantly improved by Dempster-Shafer based fusion of the single view ground occupancy maps. Experiments and a comparison with the state-of-the-art show clear improvements in the fused ground occupancy maps in terms of concentration of the occupancy evidence around ground truth person positions.

References

[1] D. Delannay, N. Danhier, and C. D. Vleeschouwer, "Detection and recognition of sports(wo)men from multiple views," in *Proceedings of ACM/IEEE ICDSC*, Como, Italy, August 2009, pp. 1–7.

[2] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 30, no. 2, pp. 267–282, 2008.

[3] A. Alahi, Y. Boursier, L. Jacques, and P. Vandergheynst, "Sport players detection and tracking with a mixed network of planar and omnidirectional cameras," in *Proceedings of ACM/IEEE ICDSC*, Como, Italy, August 2009, pp. 1–8.

[4] "Basket ball dataset from the European project APIDIS," http://www.apidis.org/Dataset/, http://www.apidis.org/ICDSC09/.

[5] A. P. Dempster, "A generalization of Bayesian inference," *Journal of the Royal Statistical Society, Series B*, vol. 30, pp. 205–247, 1968.