Visibility Management for Object Tracking in the Context of a Fisheye Camera Network

Franck Vandewiele*, Cina Motamed*, and Tarek Yahiaoui†
*Laboratoire d'Informatique Signal et Image de la Côte d'Opale
Université du Littoral Côte d'Opale, Calais, France
email: {vandewiele, motamed}@lisic.univ-littoral.fr
†Laboratoire d'Informatique Fondamentale de Lille
Université des Sciences et Technologies de Lille, France
email: tarek.yahiaoui@lifl.fr

***Abstract*—Fisheye lenses provide a wide field of view of the scene that can help reduce the number of cameras used to cover a large area. However, the geometric deformations induced by such devices can make the image interpretation difficult. For this reason, the presence of many obstacles, which is a difficult problem in every computer vision system, is especially impairing in a fisheye context, notably in the peripheral areas of the image where object visibility is already low. In this paper, we describe several generic strategies to deal with partially visible objects in a highly cluttered scene. We explain how a global model of the static obstacles can help in dealing with such problems. We present an implementation of our ideas in a real-time system backed by a network of fisheye cameras tracking customers in a retail store.**

I. Introduction

The monitoring of large indoor or outdoor areas with a person tracking system is a difficult computer vision problem for several reasons. The central reason is that it is difficult to cover hundreds of square meters with one single camera. From this simple fact emerge many problems, which largely depend on the chosen solution.

One solution is to use multiple cameras to cover a wider field of view, each covering a small portion of the area to be monitored. This popular solution introduces its own difficulties [1]. A precise calibration step is necessary to be able to express the informations emanating from several sources in the same referential. If the fields of view of the cameras happen to overlap, a measure selection step is necessary to avoid that redundant information may pollute the tracking process.

Another solution to address the problem of covering large areas is to use wide-angle lenses. The use of fisheye cameras rather than conventional perspective cameras is an inexpensive way to monitor large areas with fewer cameras. However, the distorted perception this kind of devices offer can be challenging for computer vision algorithms. One particular issue is the increased sensitivity to occlusions compared to perspective cameras. Figure 1 illustrates this problem: while a wide field of view is an attractive feature, it induces a lower robustness to partial occlusion, especially in distant areas, rendering the expansion of the field of view difficult to exploit in practice if the scene to monitor is cluttered.

To circumvent this kind of difficulty, several approaches have been proposed. Some approaches, like [2] work directly in the image plane, offering extra flexibility to the appearance model of the object tracked when an occlusion is suspected.

Some approaches rely on a precise modeling of the static obstacles.

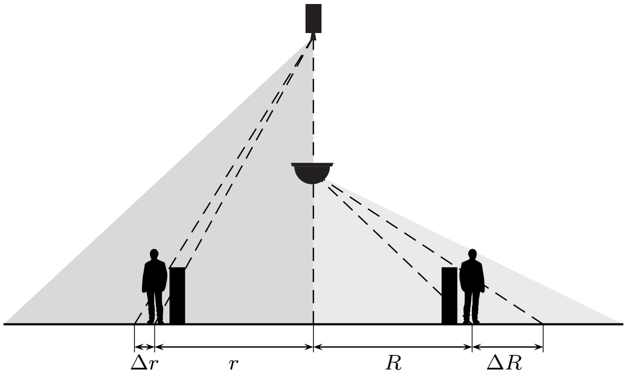

Fig. 1. Partially occluded persons perceived by a pinhole camera (left) and a fisheye camera (right) covering the same area of the ground plane. The error on the perceived position of a partially occluded person standing near the limit of the field of view of a fisheye camera is much bigger than in the case of a pinhole camera.

In [3], the authors propose to build a graphical model of the topology of the scene from occluders. In this model, an object's motion is represented as a Markov process over zones attached to occluders. This approach is yet to be extended to several overlapping cameras, though.

In [4], the authors present an automatic method to learn depth maps from large amounts of video data. They assume that pedestrians cover the whole scene and infer the structure of the scene from which pixels are occluded by a moving pedestrian and which are not, building iteratively depths maps that get finer and finer as pedestrians move across the scene. This approach suppose that such large amounts of data is readily available, and that persons moving around the scene actually cover the whole topology to estimate the depth of every occluders, which may not always always be the case in

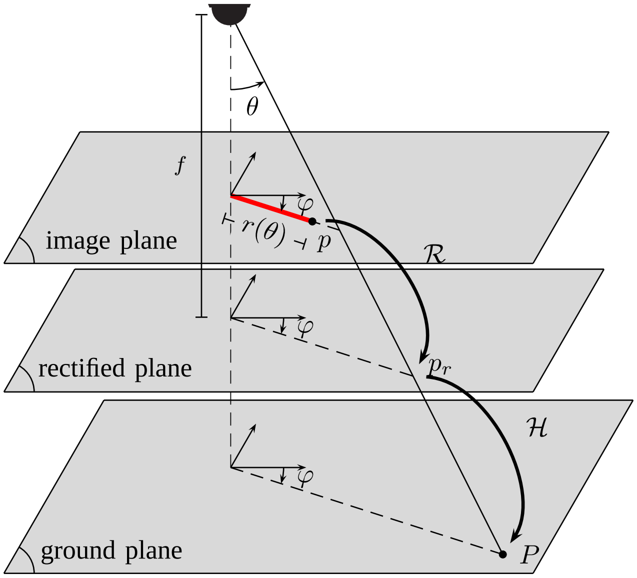

Fig. 2. Correspondences between the ground plane, the rectified image plane and the image plane using a generic camera model for fisheye image rectification. The transformation $\mathcal{H} \circ \mathcal{R}$ maps points of the image plane to the ground plane.

practice.

The authors of [5] use the notion of *visibility map* to model whether a sentry can perceive a given point in the ground plane despite the obstacles that clutter the scene. In their setup, the occluders are relocatable walls that totally occluding. No partial occluding is therefore considered.

We will present a system based on a network of fisheye cameras designed to track customers in a retail store. This kind of environment is generally highly encumbered by shelves, tables and various pieces of furniture used to display products. In our setup, the occluders are at most 1.2 meters high, so that at least partial visibility of human targets can be achieved with a reduced number of cameras. We take advantage of this partial visibility to keep the ground plane projection consistent even in zones that cannot be directly perceived by the cameras. To do so, we introduce a two-level scene modelization that grasps enough of the scene constraints to maintain tracking over time.

The rest of this paper is organized as follows. In section II, we will present the camera calibration process, along with the distortion model used to rectify fisheye images. Section III is dedicated to the presentation of our contribution: a two level modelization of the scene from known occluders. Section IV presents the tracking process and the how our modelization is used to perform data association and sensor selection. Section V presents the experimental setup and some results. Finally, some conclusions and future perspectives will be presented in section VI.

II. Camera Calibration

A. Distortion Model

Fisheye lenses cannot be modelled accurately with a modified perspective projection model. We use the generic camera model presented in [6] to represent fisheye lenses. In this model, if θ is the angle between the optical axis and an incoming ray (see figure 2), the distance r between the corresponding image point and the principal point is given by:

$$r(\theta) = \sum_{i=1}^{n} k_i \theta^{2i-1}$$

We use a two parameters model, that is:

$$r(\theta) = k_1\theta + k_2\theta^3$$

We will not take radial or tangential distortions into account, because their impact is minimal on low resolution images we will work with (240 × 240 pixels).

The inverse of this transformation, denoted $\mathcal{R}$, maps points from the image plane to points of a rectified image plane, as it would be perceived by a pinhole camera. This transformation will allow us to reason as if we were working with a perspective camera.

B. Homography-based calibration

In order to track an object in the scene, we will map points in the image plane to a ground plane. Working in the ground plane serves two purposes. First, it is a common space in which we can project information from different cameras. Secondly, it is in this plane that the topological constraints of the scene can be most easily expressed.

For each camera, there is a homographic correspondence between the rectified image plane and the ground plane. That is, there exists a 3×3 matrix H such that, for every point P lying in the ground plane and every corresponding point p in the rectified image plane, if $P = (X, Y, Z)$ and $p = (x, y, 1)$ in homogeneous coordinates:

$$P = Hp \tag{1}$$

To compute H explicitly, correspondences between salient ground plane points (P_i) and rectified image points (p_i) are established. Substitution of theses correspondences in equation 1 leads to a system whose solution is the coefficients of H. Potential outliers due to imperfect measurements are filtered out using RANSAC [7]. We denote $\mathcal{H}$ the transformation that maps points of the rectified plane to the ground plane through $P = Hp$.

C. Positioning a person in the ground plane

To position in the ground plane a target perceived in the image plane, we will use the transformation $\mathcal{H} \circ \mathcal{R}$ to map to the ground plane the point of the image plane corresponding to the contact zone between that target and the ground. Our algorithms will use the fact that, in the image plane, for a person, this contact zone (the person's feet) is the point of that person's blob that is the closest to the center of the image, as shown in figure 5.

If a person is partially occluded and the camera cannot perceive its contact zone with the ground, figure 3 shows that one can use the perceived position of that person's head in the rectified image to compute an estimate of the position of the contact zone. This transformation, which we will subsequently refer to as *height correction*, requires that the height of the human target can be estimated correctly. The height estimation process we rely on is described in section IV.

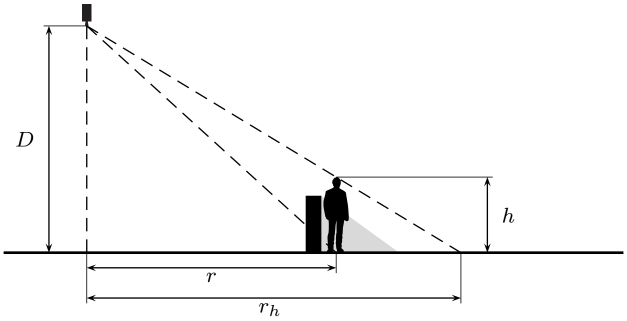

Fig. 3. While the feet of a person standing behind an obstacle cannot be perceived, their radial coordinate in the rectified image can be estimated using the radial coordinate of the person's perceived head: $r = r_h \frac{h}{D}$.

III. Global and local scene models

The scene, as perceived by our multi-camera system, is modeled with several visibility maps:

- Several *local visibility maps*, or *local scene models*, one for each camera.
- One *global visibility map*, or *global model*.

In regards to a specific camera, a given point in the ground plane can be:

- *completely visible*, that is, this point is in the field of view of the camera and not behind any known occluders. Perception of this point is perfect.
- *indirectly visible*, that is, a tall enough target standing at that point is partially occluded by a known occluder. Thus, this target's position can be computed through height correction and correctly identified to be this point precisely. In figure 3, the points of the ground plane lying in the gray triangle are indirectly visible. Perception of this point is indirect.
- *not visible at all*, because this point is under a known occluder, or behind an occluder so high that any known target would be totally occluded if standing precisely at that point. Perception of this point is impossible.

The *Local scene model* of a camera:

- keeps track of a simplified 3D model of every known occluder in its field of view;
- keeps track of which points of the ground plane are completely visible, indirectly visible through height correction and not visible at all to the camera;
- keeps track, for every indirectly visible point of the ground plane, of which obstacle model is susceptible to impair its visibility.
- maintains an *occlusion boundaries map* in the image space, computed from the 3D models of the known occluders. This map will be used during the tracking process to detect if an object may be occluded or not.

The *global scene model*:

- keeps track of a simplified 3D model of every known occluder in the scene;
- keeps track, for each point of the ground plane, of the visibility of this point for each camera.

The local scene model is used by the camera during the detection step to evaluate the reliability of its detections and perform corrections should an occlusion be detected.

The global scene model is used during the sensor selection step to filter poor detections out and reinforce the confidence in the best detections.

Both models share simplified 3D models of the known occluders of the scene which are obtained through a supervision, after a precise measurement of the building holding the scene has been performed.

IV. Tracking

A. Detection

Moving objects are detected using the classical approach of background subtraction. We relied on the method of [8] which models each pixel of the image plane with a mixture of gaussians which is updated over time. Pixels that stray too far from the model are classified as foregroud; others are classified as background.

Connected components of foreground pixels are considered to be moving objects. Isolated foreground pixels or small connected components of foreground pixels are treated as noise and removed. Because our objective is to track persons [9], we fit an ellipse around each connected component using the algorithm from [10], as depicted in figure 5. This coarse geometric model of the shape of a connected component is interesting enough at the resolutions we work with to help with several issues:

- It helps in finding the best candidates for the feet and head position of a person. We use the intersections of the ellipse with a ray going from the center of the image through the center of the ellipse to estimate the position of these important points.
- It helps to decide if the shape of a connected component is compatible with the silhouette of a human body. In peripheral zones, the orientation of the minor and major axes of the ellipse and their relative size give us important clue about the global shape of the perceives object. In these areas, the direction of the major axis should be nearly radial and the direction of the minor axis should be nearly tangential. If the principal axis of the fitted ellipse deviates too far from the radial direction, we can assume that the detection is failing or that the connected

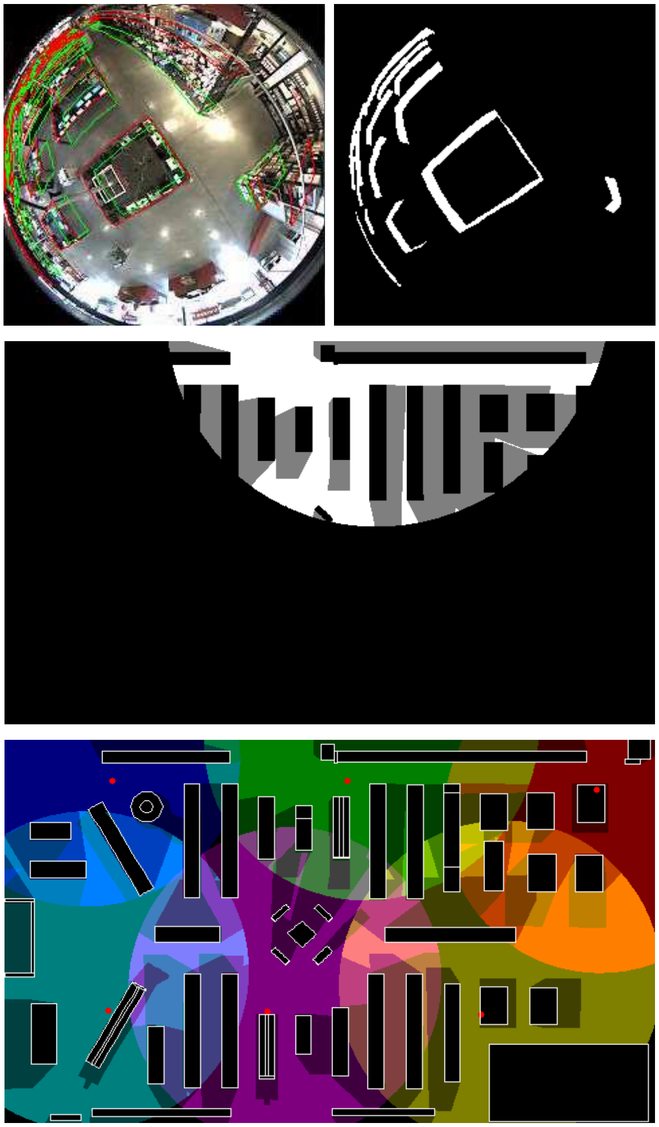

Fig. 4. Local and global scene models. Top left: image perceived by a camera, enhanced by a representation of the known occluders. Top right: corresponding occlusion boundaries map. Center: a representation of the local scene model as perceived by one of the cameras. Black areas are *not visible*, because impassable or very distant: positioning in these zones is considered unreliable; gray areas are *indirectly visible*: objects standing in these zones are partially occluded and their positioning is considered reasonably reliable; white areas are *completely visible*: positioning in these zones is considered reliable. Bottom: a representation of the global scene model. Darker areas are less visible.

component actually encompasses two different targets. In central zones, this is no longer true. Viewed directly from above, the persons tend be perceived as a circular shape and the axes no longer provide any interesting informations.

Connected components that do not lie on an occlusion boundary are projected to the ground plane.

If the point closest to the center of the image of the fitted ellipse of a connected component lies on an occlusion boundary, it means that the corresponding object is possibly occluded. In that case, the point of the fitted ellipse which is the farthest from the center of the image is used as a head candidate for the target and its position in the ground plane is height corrected accordingly. The local model checks that the height correction positions the object in an indirectly visible zone of the ground plane. If it were not the case, the detection would be ruled unreliable and discarded.

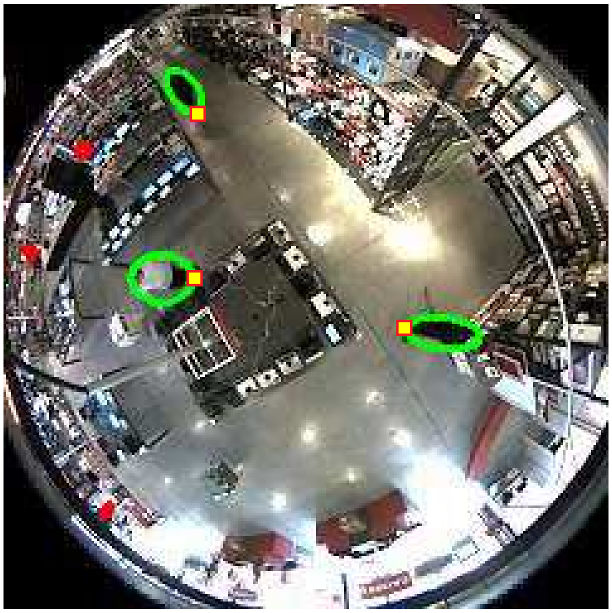

Fig. 5. Ellipses are fitted around connected components of foreground pixels. For each connected component, yellow squares mark the point of its fitted ellipse that is the closest to the center of the image and is considered to be in contact with the ground and mapped to the ground plane.

B. *Visual cues extraction*

Depending on the region in which a connected component lies in the image plane, we may extract some consistent visual cues to help with data association.

1) Region-specific extraction: Extreme peripheral areas of the image perceive connected components of very small area, only a few pixels. No consistent visual cues can be extracted from such few information.

Central areas of the image perceive objects from above. A visual appearance model specific for these areas is captured. No height information can be retrieved.

Peripheral areas of the image perceive objects from aside. If a connected component is not occluded, a visual appearance model specific for theses areas can be captured, as well as a height estimate of the object.

2) Height Estimation: Estimation of the height of a human target is essential for the height correction process. Because no height information can be retrieved from the extreme peripheral and central areas of the image, the height of a newly tracked human target appearing in these zones cannot be estimated correctly until it enters a peripheral zone of the image. We use the median of the height of the human population as a temporary height value to perform the height

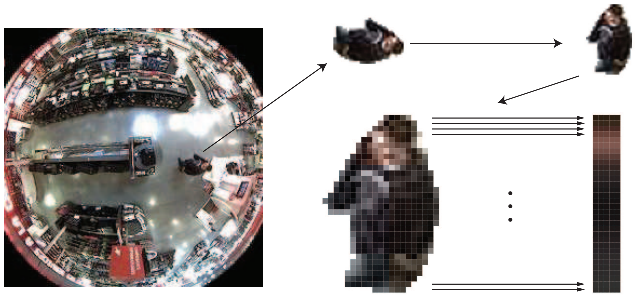

Fig. 6. Computation of a connected component's visual model. From the camera image (left), a connected component is extracted and aligned with a top-down radial direction (right). The visual model is essentially a collection of vectors, one by color channel, which elements are the mean of the color channel in the corresponding row of the aligned component (bottom). These vectors are finally resized to a normalized height for further analysis.

correction until a better estimate is available. The position in the ground plane will be corrected during the association step. When an estimate of the height of a target becomes available, the whole track is corrected with this new value of the height.

3) Appearance model extraction: Due to the intrinsic symmetry of fisheye lenses, connected components need to be aligned with the same radial direction to be compared. As depicted in figure 6, we basically slice the connected component in rows and take the mean of each row to form a descriptor of the color distribution in the image. This computation is made in the Lab color space to ensure color information consistency between different cameras.

C. Data association and Sensor selection

In the areas where cameras overlap, it is important to select which sensor provides the most reliable information to avoid poor data association or the generation of additional tracks if redundant detections are not filtered out.

When several detections are positioned close to each other, the global scene model decides which camera provides the most reliable information, based on the visibility of the centroid of the projections in the ground plane of the components susceptible to append a given track.

Once association is performed, each height corrected detection is adjusted with its associated track's height estimate, part of the appearance model, if the track happens to have one instanciated.

Each track has its own appearance model, consisting of the average and variance of the detections it is associated to. When the association is completed, each track's appearance model can be updated with its newly associated detection's own appearance model, using recursive average and variance computation methods.

V. Experimental setup and evaluation

Our system consists of six Mobotix Q24 fisheye cameras and a central server powered by an Intel Core-i7 processor. We implemented our algorithms in C++ using the OpenCV [11] library.

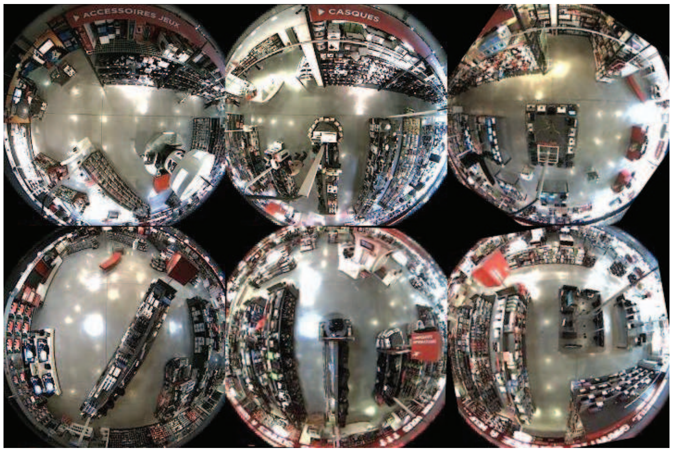

Fig. 7. The scene, as it is seen by the six fisheye cameras of our system. Due to the high number of occluders and the low resolution (240 × 240 pixels) of each camera view, distant zones are not easily perceived.

	T1	T2	T3
TMEMT	31.82	13.27	7.00
TMEMTD	33.92	18.20	9.28
TF	11.21	3.41	2.83

TABLE I
Results of the series of tests T1 (pure Kalman filter), T2 (Kalman filter + height correction), and T3 (whole models of the scene).

To evaluate the performance of our approach, we used several classical evaluation metrics for object tracking evaluation, as described in [12] in different conditions. We ran a series of tests on a sequence of 1100 frames. In this sequence, 14 ground truth tracks were compared to the tracks the system perceived.

We performed a series of three different tests. In the first series, our whole scene model was inactive, and a simple Kalman filter tracking was performed. In the second series of tests, the height correction process of our system was the only component activated. Sensor selection was performed assuming that the closer a sensor is to a point of the ground plane, the better the information it delivers, regardless of the occluders topology. In the third series, we used the complete system. Results are depicted in table I.

We used three metrics to perform our tests. Track Matching Error (TMEMT) and Track Matching Error Standard Deviation (TMEMTD) measures the positioning error of the system. Track Fragmentation (TF) is a measure of discontinuity of the tracks perceived by the system.

VI. Conclusion

We have presented a complete person tracking system using a network of fisheye cameras based on a two-level modelization of the static obstacles of the scene. The local model proves itself useful during the detection step to ensure that the positioning of detected objects in the ground plane takes a possible occlusion into account. The global model is

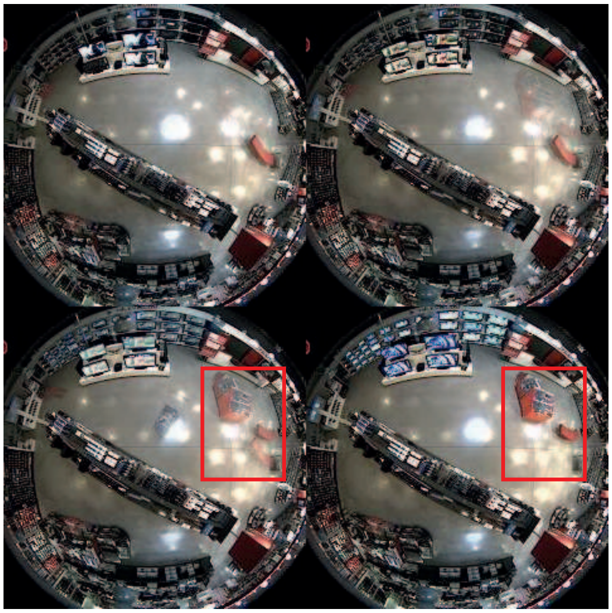

Fig. 8. A relocatable occluder is repositioned in the scene. A slow-paced background subtraction method can perceive the topological change without raising false alerts.

an interesting tool during the tracking process to make the best decision for sensor selection.

Experimental results show that our approach offers a significant performance increase on a difficult dataset compared to less robust methods. The tracks generated by our system are much less fragmented, which helps preserving the target identity during the tracking process.

Our approach relies on a supervised modelization of the obstacles of the scene using simplified 3D models. The system itself cannot detect and integrate new obstacles by itself. Obviously, this is not completely satisfactory. In the context we are working in, many occluders are actually relocatable, and our modelization is not robust to such changes for the moment. We are currently experimenting a long-term background subtraction system, that operates in parallel of the detection process, but which updates the background changes more slowly to capture important changes in the topology of the scene. Figure 8 shows a topological change detected by that strategy. Currently, the system raises an alert when a possible topology change is detected, but does not updates itself. One idea of [3] could be to build a database of possible occluders and try to estimate which corresponds best to the topological change.

Acknowledgements

The authors would like to thank the ANAXA$_{\text{VIDA}}$ company [13] for providing the data used in this work and providing access to their hardware and prior knowledge of the scene.

References

[1] H. Aghajan and A. Cavallaro, *Multi-Camera Networks: Principles and Applications*. Academic Press, 2009.

[2] C. Zhang, J. Xu, A. Beaugendre, and S. Goto, "A klt-based approach for occlusion handling in human tracking," in *Picture Coding Symposium*, 2012, pp. 337–340.

[3] V. Ablavsky and S. Sclaroff, "Layered graphical models for tracking partially occluded objects." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1758–1775, 2011.

[4] D. Greenhill, J. Renno, J. Orwell, and G. A. Jones, "Occlusion analysis: Learning and utilising depth maps in object tracking," in *In Proceedings of British Machine Vision Conference*, 2004, pp. 467–476.

[5] P. Chakravarty, D. Rawlinson, and R. Jarvis, "Covert behaviour detection in a collaborative surveillance system," in *Proceedings of the Australasian Conference on Robotics and Automation*, Dec. 2011.

[6] J. Kannala and S. Brandt, "A generic camera calibration method for fish-eye lenses," *Pattern Recognition, International Conference on*, pp. 10–13, 2004.

[7] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.

[8] Z. Zivkovic, "Improved adaptive gausian mixture model for background subtraction," in *Proceedings of the International Conference on Pattern Recognition*, 2004.

[9] Y. Kubo, T. Kitaguchi, and J. Yamaguchi, "Covert behaviour detection in a collaborative surveillance system," in *Proceedings of SICE Annual Conference*, Aug. 2011, pp. 2013–2017.

[10] A. Fitzgibbon and R. B. Fisher, "A buyer's guide to conic fitting," in *British Machine Vision Conference*, 1995, pp. 513–522.

[11] G. Bradski and A. Kaehler, *Learning OpenCV*. O'Reilly Media Inc., 2008.

[12] F. Yin, D. Makris, and S. A. Velastin, "Performance Evaluation of Object Tracking Algorithms," in *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2007)*, Oct. 2007.

[13] ANAXA$_{\text{VIDA}}$ company, http://www.anaxa-vida.com.