

People Detection and Tracking with World-Z Map from a Single Stereo Camera

Sofiane Yous
Trinity College, Dublin
s.yous@ieee.org

Hamid Laga
Tokyo Inst. of Technology, Japan
hamid@img.cs.titech.ac.jp

Kunihiro Chihara
NAIST, Japan
chihara@is.naist.jp

Abstract

In this paper we propose a new people tracking system that uses a single stereo camera fixed at a high position and observing the scene at an oblique angle. We introduce the notion of world-Z map and show that 3D people detection and segmentation can be performed efficiently in this map and outperforms significantly the methods based on depth or plan-view. Detection and tracking play complementary roles, we derive a probabilistic framework for tracking based on motion. The system successfully deals with very complex situations without losing track of people in highly cluttered scenes for long observation periods and achieves around 10 fps on a single PC.

1 Introduction

People tracking is a very active research field that aims at detecting people within a monitored space and keeping tracking them as long as they are inside. The importance of people tracking is related to the range of applications that can take benefits from it such as visual surveillance and people counting. It is needless to argue how visual surveillance is important nowadays. Cameras are more and more deployed in indoor and outdoor public spaces, such as shopping malls, retails and parking lots, for monitoring and helping the detection of abnormalities. They have been recently used not only for security purposes but also to draw some statistics useful for planning and management.

The purpose of people tracking is to provide a set of tracks that are in a one-to-one correspondence with people appearing in the monitored area [13]. The ideal system should satisfy the following requirements:

- *Tracking in complex situations* such as crowded scenes where occlusions are frequent. Previous work [13] identified three types of occlusions; (1) *partial occlusions* where only a part of the subject is visible to the sensor, (2) *short-term occlusions* where the subject is completely occluded for short periods of time, and (3)

extended occlusions where the subject leaves the field of view for an extended period.

- *Robustness against lighting changes and background complexity* is a major concern. Robust systems should be able to operate in indoor and outdoor environments, and handle scenes with complex and dynamic backgrounds.
- *Accuracy* in recovering the trajectory of each individual with respect to the 3D space or floor plane.
- *Fast processing* in applications such as visual surveillance where the reaction time is critical. In these situations, the system should provide real-time information required for behavior analysis and abnormalities detection.
- *Low deployment cost and easy setup and maintenance.* Ideally the system should have minimum free parameters to tune.

Although many solutions have been proposed in the literature, none of them is able to fulfill simultaneously all the requirements. In this paper we propose a new multi-people tracking system that uses a single stereo camera fixed at a high position and observing the scene at an oblique angle. We introduce the notion of world-Z map and show that people detection and segmentation can be performed efficiently outperforming significantly the existing methods in general and the plan-view-based methods [11] in particular (Section 2.1). The detection and segmentation process is described in Section 2.2. For tracking, we derive a probabilistic framework which allows the tracking of variable number of people in real-time (Section 3). We evaluated the system in very complex situations (Section 4) where the results show that it outperforms significantly the state of the art methods. Section 5 summarizes the main contributions of the paper and outlines some issues for future work.

1.1 Related work

There is a large amount of research on people tracking as it rises many challenging issues to image processing, com-

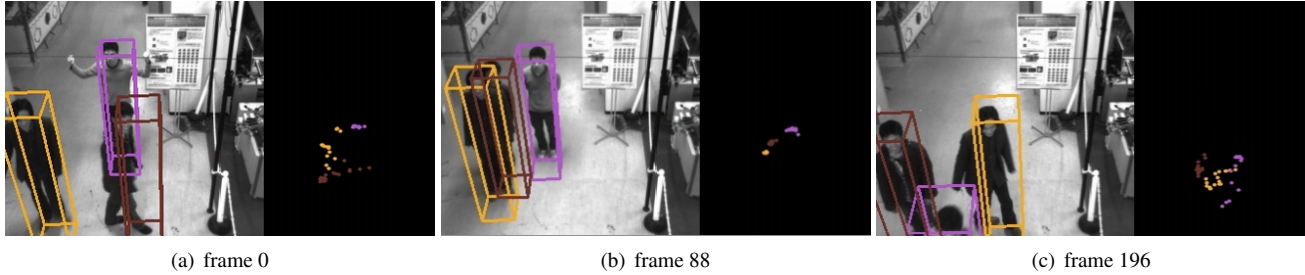


Figure 1. Stereo-based 3D people tracking: the system deals with complex situations.

puter vision and machine learning communities. Dealing with occlusions and lighting changes in multi-person tracking are among the most challenging ones.

Early works used a single monocular camera [14, 15]. Since a single monocular camera does not provide 3D information of scene, dealing with partial and short-term occlusions requires strong prior knowledge about the geometry, appearance and dynamics of the objects to track. This knowledge is used to build a Bayesian framework for detection and tracking [14, 15]. However, strong priors limit the flexibility of the system to handle unseen objects with high intra-class variability.

Recent advances in computing and imaging hardware had motivated the use of distributed monocular camera systems to gain access to the 3D information of the scene. Chai and Aggarwal [8] developed a distributed camera system for motion tracking in structured environments. Fleuret et al. [11] used two to four synchronized video streams taken at eye level from different angles to detect and track up to six people in the presence of occlusions. They use a *probabilistic occupancy map* which can be affected by: 1) the poor quality of background subtraction, 2) the presence on people in an area covered by only one camera and 3) the excessive proximity of several people. Their multi-person tracking is achieved by processing individual trajectories separately over patches of 100 frames.

Stereo cameras can be seen as a particular case of distributed monocular camera systems that produce dense depth maps of the scene. Augmenting the input space with depth maps provides better robustness to many algorithms including background subtraction, and object detection and tracking. Stereo-based object detection and tracking has been previously used in stationary rig setup [7, 9], and mobile platforms [6]. These works exploit the fact that stereo provides a stable basis for object segmentation, resolves to some extent the occlusions and provides useful 3D cues to the tracking stages, such as 3D positions and relative object sizes [7]. Harville and Li [12] and Darell et al. [9] developed a powerful tool to solve the problem of partial and short term occlusions. They use the depth maps

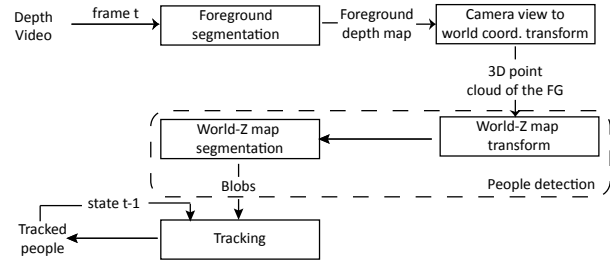


Figure 2. Overview of 3D people tracking.

to estimate a virtual overhead view, called the plan-view, and perform the detection and tracking in this plane based on some statistics, such as height and occupancy, recorded on this plane. For instance, Harville and Li [12] combine occupancy and height maps to eliminate the 3D noise and deal with occlusion.

Although plan-view-based methods handle many occlusions, it is a lossy representation in the sense that the rendered top view keeps only the top of objects present in the scene. This fact limits the processing that can be done on this plane. In this paper, we propose a new map, called *world-Z map*, that records the world-Z coordinates of the scene points into the image plane. In addition to the fact that this map has the advantage of being lossless, since it represents every point in the original depth map, we will show in Section 2.1 that it underlies very useful properties that enable high-performance segmentation and, hence, efficient and accurate detection.

1.2 Contributions

In this paper, we propose a new multi-person tracking system that deals the best with very complex situations. Specifically we make the following contributions:

- We introduce the concept of world-Z map and show that people detection and segmentation can be efficiently performed in this new plane compared to methods based on depth and plan-view.

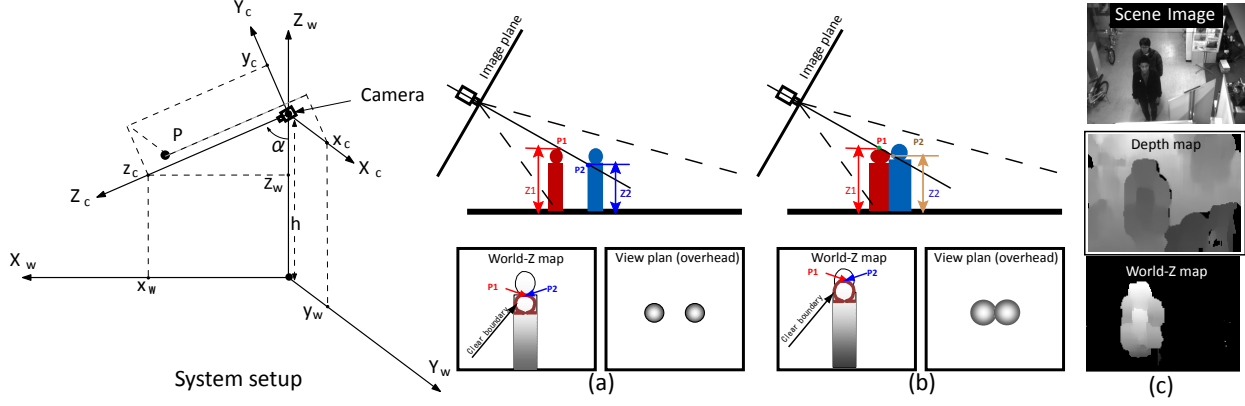


Figure 3. Building the world-Z map: (a) when the two objects are clearly separated both the plan-view and world-Z map allow efficient segmentation. (b) The world-Z map maintains clear boundaries between very close and partially occluded objects. (c) world-Z map of a scene with two persons.

- The system we propose deals efficiently with partial and short term occlusions and computes accurate 3D trajectories of the individuals.
- Only the height and viewing angle with respect to the floor plane are required. We show that this is a very easy task and therefore, our method does not require a complex calibration step.
- The entire detection and tracking process performs in real-time. In our experiments the lowest frame-rate was 10fps on scenes of nine people.

The system operates in an online mode and, therefore, is well suited for real-time applications such as visual surveillance, people counting and crowd monitoring.

2 3D people detection

To address people detection challenge, depth video cameras are the most suitable. Such cameras provide an image of the scene where each pixel contains, in addition to the intensity, the distance to the subject with respect to the camera, called depth. There are many commercial cameras generating depth information at video rates [2, 3, 4, 5], and the technology is evolving rapidly. We expect that prices will continue dropping and depth video cameras will be available at reasonable costs in near future.

Using depth information, previous works compute the 3D data and map them onto an overhead virtual plane called plan-view [9, 12]. Plan-view can record many useful statistics, such as occupancy [11, 9] and people height [12]. Although interesting results have been achieved, the segmentation is not efficient when the scene contains adjacent objects touching each other, which occurs often in crowded

scenes. Furthermore, the occupancy and height maps are lossy in the sense that not all 3D points are represented.

In this paper we propose a new map Φ on which the detection is efficient even in very crowded scenes. First we will describe the new map and then we will show how we perform people detection.

2.1 World-Z map

The world-Z map records on the image plane the world-Z coordinates of the 3D points of the scene. Let's consider that the camera is set at a height h from the floor and observing the floor at an oblique angle α . We assume also that the XY plane is parallel to the floor and the coordinates of the camera center are $(0, 0, h)$, see Fig. 3. Given a point p with camera coordinates (x_c, y_c, z_c) , its world coordinates (x_w, y_w, z_w) are given by:

$$x_w = z_c \sin(\alpha), \quad y_w = x_c, \quad z_w = h - z_c \cos(\alpha). \quad (1)$$

The world-Z map records on the image plane, the z_w coordinates of every foreground point. This representation is lossless since all 3D points computed from the initial depth map are represented. By setting the stereo camera at a high position looking towards the scene at an oblique angle with the floor, the points belonging to the closer objects with respect to the camera have larger attributes on the world-Z map, creating an implicit boundary between objects. This property is very important since it enables easier segmentation. When using the depth map, the boundaries between objects can be detected using bidirectional threshold which is not easy to set. When using the world-Z map, however, no threshold is needed as will be explained in section 2.2.2.

Fig. 3(c) shows the world-Z map computed from a scene containing two individuals. We will see in the next section

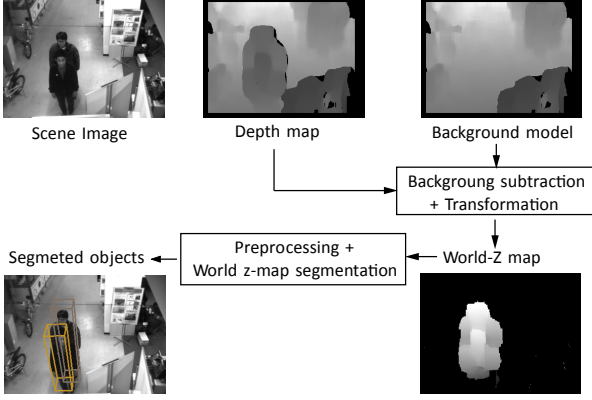


Figure 4. 3D segmentation-based foreground objects detection

the properties of this map and how they are useful for segmentation of objects in general and people in particular.

2.2 People detection

In this section, we formulate the people detection problem as a world-Z map segmentation. Prior to building the world-Z map, we apply a depth-based background subtraction followed by a preprocessing step. The entire process is illustrated in Fig. 4.

2.2.1 Preprocessing

The preprocessing step consists of eliminating the noise inherent in stereo-based depth estimation. These errors can create false connections between occluding and partially occluded objects and affect the segmentation performance. The main noise that affects segmentation is located at the object boundaries. It is mainly due to the windowing used in stereo processing. Harville and Li [12] eliminate this noise using the occupancy map. However, this method fails when the noise belongs to occupied bins.

The noisy points are characterized by a large angle of incidence with respect to the camera, see Fig. 5(a). We use this property to setup a thresholding criteria to filter this noise. We build an incidence map in the image plane and record at each pixel the angle of incidence of a point in the point set. Next, we filter this map to keep only the points having angles of incidence smaller than a predefined threshold. This map will be used as a mask for the world-Z map.

The cosine of the angle of incidence at a given point p is the dot product between the unit normal vector \vec{n}_p at p and the unit vector \vec{v}_p connecting the camera center c to p . The incidence map entry $I(p)$ corresponding to the point p is given by $I(p) = \arccos(\vec{v}_p \cdot \vec{n}_p)$, where \vec{n}_p is computed

using two other valid neighbors p_1 and p_2 not aligned with p , and is given by the normalized cross product between $\overrightarrow{pp_1}$ and $\overrightarrow{pp_2}$.

Fig. 5(b) shows the effect of the preprocessing step on eliminating the noise and preparing the data to the next step.

2.2.2 Using the world-Z map for people detection

In this section we make use of the properties of the world-Z map to design a powerful algorithm for object segmentation that deals the best with partial and short-term occlusions including situations where objects are very close or touching each other. Particularly:

1. The world-z map is a lossless representation as compared to the plan-view.
2. The points of an object (one connected component) are always ordered inversely to the Y axis of the camera.
3. Points belonging to closer objects with respect to the camera have higher values on this map. This means that if the objects in the scene have roughly similar size, as it is the case of people, the most top point in the world-Z map belongs to the farthest object in the scene with respect to the camera.

To segment the back object from the front ones (all other that are closer to the camera), we start from the top-left point $p(x, y)$ of the World-Z map and scan the image top-down and left to right to find neighbor points $p'(x', y')$; p and p' belong to the same segment if and only if $\Phi(x', y') + \epsilon \leq \Phi(x, y)$, where ϵ is a user defined threshold. Based on this, people segmentation can be summarized as follows:

1. Scan the world-Z map from top to down and from left to right. Exit when the last point is reached.
2. Recursively find, in the same scanning direction, all neighbors that have lower values in the world-Z map.
3. Remove the segmented region from the world-Z map and go to step 1.

For comparison, recall that the condition for the segmentation of the depth map D is: $D(x', y) - \theta_l \leq D(x, y) \leq D(x', y) + \theta_u$, which requires the setting of two thresholds θ_l and θ_u .

Fig. 4 demonstrates how well the segmentation method deals with complex occlusions. When two persons are at the same distance from the camera and touching each other, the lower part of one object will be detected as belonging to the object segmented first. For tasks, such as people counting, this is sufficient since both objects have been detected. Such issue can be also easily solved by limiting the search space during the segmentation process. The method also

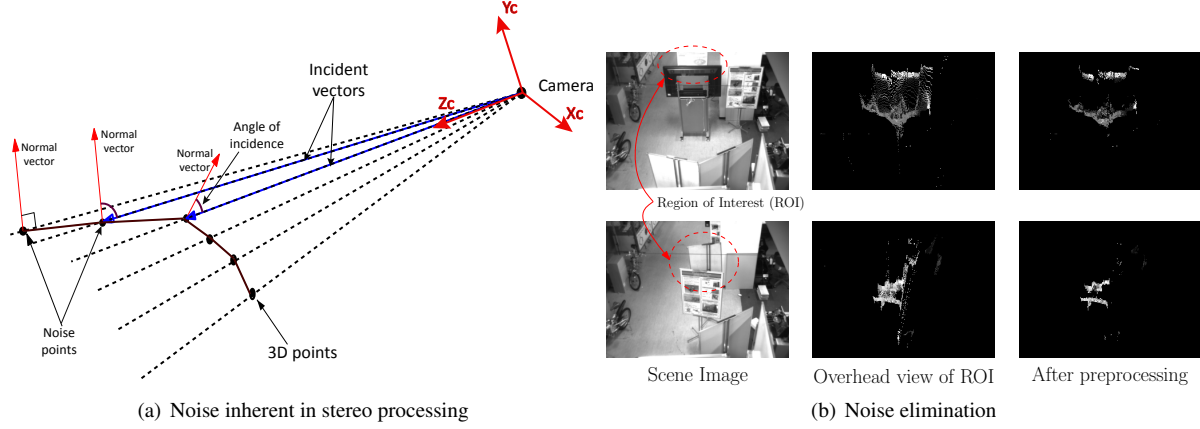


Figure 5. Preprocessing: elimination of the noise inherent to stereo processing.

can detect some false positives such as the case where an individual raises his hand. This will be resolved by the persistence test in the tracking step.

3 People tracking

Our goal is to track unknown number of people using the world-Z map as our observations. In this section we formulate the problem and discuss the implementation details.

3.1 Problem formulation

First, we assume that the number of people to track can vary with time. This allows the tracking of new objects entering the monitored area as well as handling situations where the track of some objects is lost. This may occur if an object exits the scene, becomes occluded by another object, or missed by the detection algorithm.

A multi-person tracking problem consists of establishing, at each time step t , an association between the state vector $X^{t-1} = (x_1^{t-1}, \dots, x_{m_{t-1}}^{t-1})$ at time $t-1$ with the current measurements $Y^t = (y_1^t, \dots, y_{n_t}^t)$. The component x_i^t is the state corresponding to the i -th tracked person at time t . It encodes some of its features that are relevant for tracking. This can include position, geometry, appearance and motion. Harville and Li [12] defines x_i^t as the estimated position, velocity and body configuration of the i -th person. In our implementation we use the estimated position.

The measurements Y^t at time t are obtained using the object detector algorithm described in Section 2.2. To handle the cases of *lost* objects and tracking of *new* arrivals, we set $n_t = k_t + m_{t-1}$, where:

- $k_t, k_t \geq 0$, is the number of blobs observed at time t .

- $y_{k_t+l}, l = 1, \dots, m_{t-1}$, sometimes denoted by the additional variables ϕ_l for clarity, is the observation corresponding to the case where the l -th object x_l^{t-1} detected at time $t-1$ is lost at time t .

The tracking problem can now be reduced to the problem of finding the configuration X^t that maximizes the posterior probability $P(X^t|Y^t)$. The solution is given by:

$$\hat{X}^t = \arg \max_{X^t} P(Y^t|X^t)P(X^t|X^{t-1}). \quad (2)$$

Equation 2 assumes that the current state depends only on the previous one. The algorithm that finds the optimal solution can be summarized in three steps:

- **Step 1 - Initialization:** the system remains idle until the first objects are detected. These will be the initial set of the observation vector Y^0 and state vector X^0 . At this stage, the time t is set to zero.
- **Step 2 - Prediction:** we predict the new state \hat{X}^t that maximizes $P(X^t|X^{t-1})$. It is defined as:

$$\hat{X}^t = f(X^{t-1}). \quad (3)$$

where f is the prediction function called also motion model. It can be chosen to be linear such as Kalman filter [7, 9], non-linear such as the particle filter [10], or by assuming constant velocities [12]. In our implementation we assume a linear motion model.

- **Step 3 - Data association:** assuming that the random variables $x_i^t, i = 1, \dots, m_t$ are independent, we have:

$$X^t = \arg \max_{X^t} P(Y^t|\hat{X}^t) = \prod_{i=1}^{m_t} P(Y^t|\hat{x}_i^t). \quad (4)$$

We detail this step in Section 3.2.

This algorithm establishes a correspondence between the objects tracked in the previous frame and the newly detected blobs. States x_i^t assigned to any of the additional variables ϕ_l will be considered as lost. Observations y_i^t left without any assignment will be considered as new objects entering the scene and will be added to the state vector X^t .

In our implementation we assume that: (1) at time $t < 0$, the scene is empty; (2) the objects should enter the scene from its boundaries; and (3) each object entering or exiting the monitored area goes first through a transit area defined along the boundaries. In this area the objects are detected but not incorporated into the tracking loop.

The second assumption allows to solve the problem of background noise and avoid the detection of static objects in the foreground. Although it appears very restrictive, it is valid in most of video surveillance applications, where the objects to track usually come from outside the monitored area. They are only tracked once they enter the scene.

3.2 Data association

The correction step requires an exhaustive search in order to find the optimal association. We start by defining a matrix M^t of size $n_t \times m_t$, where the entry $M_{ij}^t = P(y_i^t | x_j^t)$ is the probability of the observation y_i^t conditioned on the state x_j^t estimated at the prediction step. We assume that it is a 2D Gaussian centered at x_j^t and of standard deviation σ in both directions. σ is also the radius of the search area around x_j^t , where observations y_i^t that fall inside the search circle have higher probability of being associated to the configuration x_j^t . It is closely related to the speed of movement of the objects in the scene as well as to the frame-rate of the tracking algorithm. We define σ to be twice the expected displacement per frame. In our implementation we assume that people are moving with the average speed of $v_p = 3.5\text{mph}$, and a tracking at an average rate of 15fps.

A configuration x_i should be associated to a lost state $\phi_i, i = 1, \dots, m_{t-1}$ if either the object has exited the tracked area or the detection failed because of occlusion, or a jump due to a movement speed that exceeds significantly the radius of the search area. Therefore, the conditional probabilities of the variables ϕ_l are given by:

$$P(\phi_i | x_i^t) = \left\{ 1 - 2 \int_0^\sigma \mathcal{N}(x, 0, \sigma^2) dx \right\} + \mathcal{N}(\delta(x_i^t), 0, \sigma^2) \quad (5)$$

where $\mathcal{N}(x, 0, \sigma^2)$ is the normal distribution of zero mean and variance σ^2 evaluated at point x , and $\delta(x)$ is the minimum distance of x to the boundaries of the tracking area. The first term of Equation 5 accounts for the probability that x_i^t has moved from frame $t - 1$ to frame t with a distance larger than the radius σ of the search area and can be

computed in advance. The smaller σ is the larger is this probability. The second term corresponds to the probability that x_i^t has exited the tracking area.

Equation 5 gives the exact quantification of the probability of losing an object x_i^t . In our experiments, however, we noticed that there was not any performance improvement when using simplified models such as considering only the first term, as well as considering only the probability of falling on the edge of the search area, i.e., $P(\phi_i | x_i^t) = \mathcal{N}(\sigma, 0, \sigma^2)$. We used this last model in all the experiments shown in this paper.

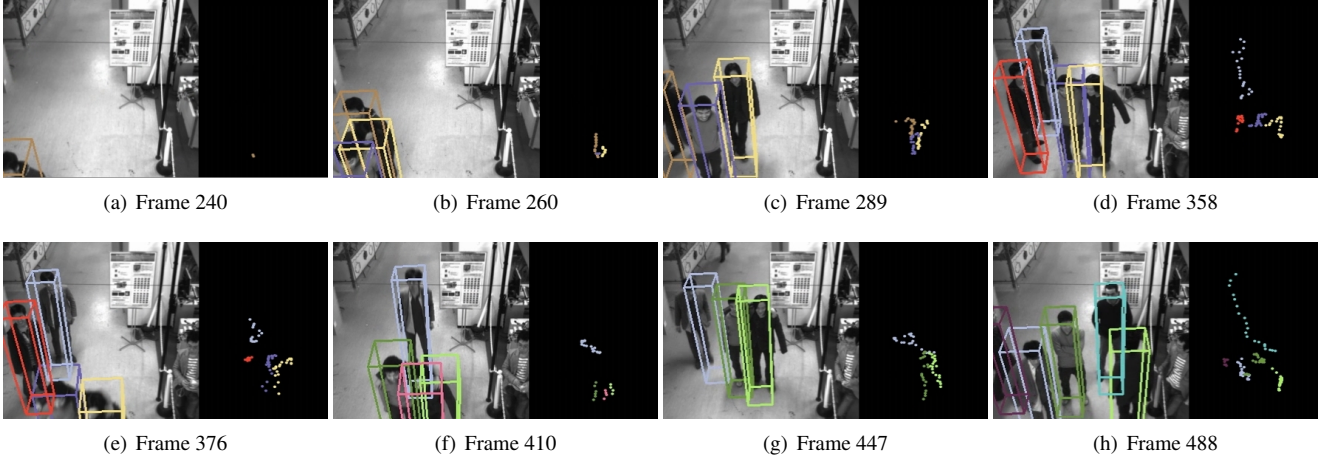
Once the conditional probabilities are defined, the matrix M^t can be fully constructed and normalized such as the columns sum to one. However M^t is a full dense matrix and an exhaustive search to find the optimal association is computationally prohibited especially when tracking a large number of objects. To overcome this problem, Fleuret et al. [11] rely on the appearance model to establish an ordering of the observations that matches the ordering of the state vector X^{t-1} . Hence, the association problem becomes convex and can be solved in a linear time. Such solution is not guaranteed to be the optimal one for the original problem. Furthermore, the appearance model introduces high probability of identity confusion and switch.

We introduce a simple but efficient heuristic that minimizes the probability of identity confusion and switch. The main idea is to make M^t sparse while maintaining the high tracking precision. We do this by setting to zero the entries where the Euclidean distance between the predicted state vector x_i^t and the observation y_j^t exceeds the radius σ of the search area. This is well justified since the Gaussian function vanishes rapidly. This allows us to achieve the tracking in crowded scenes with severe overlaps between objects at more than 10fps.

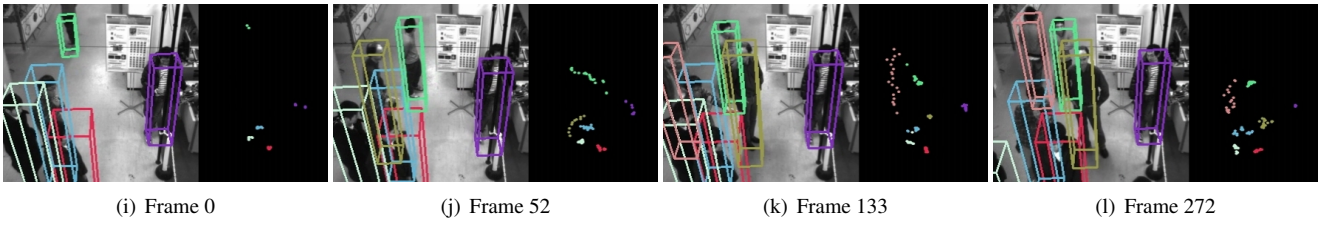
In summary, the tracking procedure has the following properties:

- Unlike previous work that uses a sliding window (of size 100 in [11] for example), the tracking at frame t depends only on the previous frame.
- False detections do not occur inside the detection area because our algorithm does not allow new detections in this area, and assumes that any new person should enter through the transition area.
- False detections may occur only inside the transition area. For example, if a person enters by the transition area with his arm up, two persons will be detected. However, because of the persistency test, the false positives will be discarded earlier in the process.

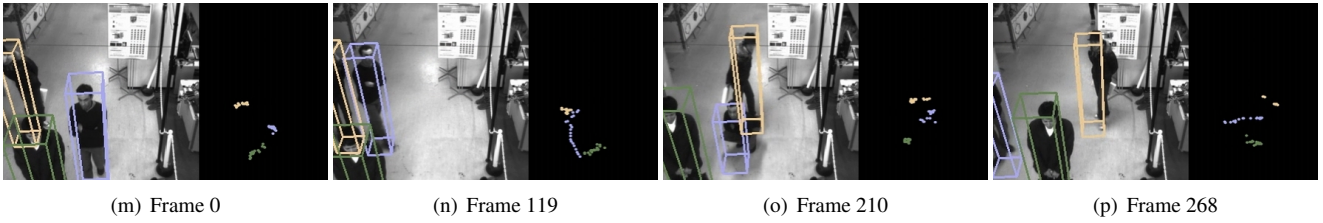
Sequence 1



Sequence 2



Sequence 3



Sequence 4

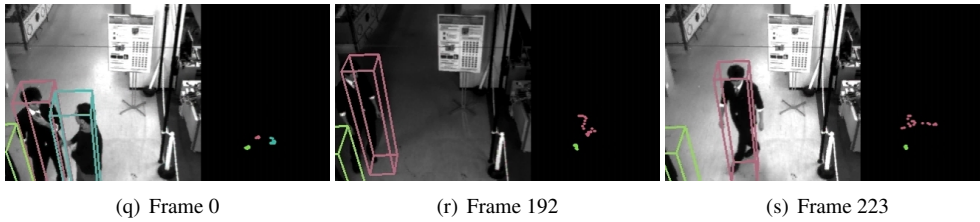


Figure 6. The performance of 3D people tracking in complex situations, see the submitted video and more examples at the project webpage (<https://www.cs.tcd.ie/Sofiane.Yous/Projects/PeopleTracking/>).

4 Results

The system was tested in real situations during an open campus event. It achieved high performance when dealing with very complex situations without losing tracks of people for long observation periods, as illustrated in Fig. 6, see

the submitted video.

The sequence in Fig.1 and Fig.6(a) to 6(h) is of extreme complexity and no previous system can deal with. It contains scenes with very high occlusions and split-merge situations. The person on the right of the Figure 6(d) to 6(h) is not detected because it is outside the overlapping area of

the stereo camera where depth is computed. We see also a false detection in Figure 6(f) but quickly discarded since it is not persistent in time. Such false detections occur usually in the transition area but rarely in the tracking area. This can be seen in frame 0 of Figure 1(a); although a raised hand is detected as an object and since no new detections are allowed in the tracking space, the tracking discarded it. The second sequence that starts at Fig. 6(i) shows a scene with many people in a limited space. Even in the presence of high occlusions, the system does not lose track of them.

The third sequence demonstrates how the system behaves when an individual sits down. Since no restrictions have been made on the height, unlike in [12], the system keeps tracking the individual. This sequence also demonstrates the robustness of our system to lighting changes. Although part of the room light was turned off the tracking has not been affected. This robustness allows the system to work indoor and outdoor without parameter tuning.

To further evaluate the performance of our system, we computed the False-Positive (FP) and False-Negative (FN) rates in a sequence with one to nine people moving freely in the scene. The estimated FP rate was 2.48% while the FN rate was 3.95%. These rates are, respectively, 3.99% and 6.14% in [11] where some constraint about number and the movement of people were imposed, See [1]. In contrast, in our test people were moving freely and were not aware of our system or trying extreme situations. The video sequences has been taken in an open campus event, see <https://www.cs.tcd.ie/Sofiane.Yous/Projects/PeopleTracking>.

The system achieves about 10fps independently of the number of individuals in the scene. Therefore, it is very suitable for applications that require real-time operation such as visual surveillance.

5 Conclusion

We proposed a new system for 3D people tracking using a single stereo camera. We proposed a new method for people detection based on 3D segmentation of the point sets obtained by stereo processing. We introduced the world-Z map which records the world-Z coordinates of the points of the scene on the image plane. The properties of this map allow accurate object segmentation under high partial and short-term occlusions. The performance of people detection allows a simple probabilistic framework to achieve high performance tracking. Our algorithm achieves a good performance in terms of accuracy, robustness to complex lighting changes, occlusions, and processing time.

As future work, we will explore whether introducing other object properties improves further the tracking performance. We plan also to investigate the suitability of the world-Z map for detecting and tracking other objects such as cars in roads and parking lots.

Acknowledgement

Hamid Laga is supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology program *Promotion of Environmental Improvement for Independence of Young Researchers* under the Special Coordination Funds for Promoting Science and Technology.

Sofiane Yous would like to thank Prof. Masatsugu Kido for his invaluable advice and unlimited support.

References

- [1] <http://cvlab.epfl.ch/projects/cti/surv/#pom>.
- [2] <http://www.3dvsystems.com/>.
- [3] <http://www.canesta.com/>.
- [4] <http://www.primesense.com>.
- [5] <http://www.vialux.de/>.
- [6] M. Agrawal, K. Konolige, and L. Iocchi. Real-time detection of independent motion using stereo. In *7th IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing (WACV/MOTION 2005)*, pages 207–214, 2005.
- [7] D. Beymer and K. Konolige. Real-time tracking of multiple people using continuous detection. *International Conference on Computer Vision (ICCV1999)*, 1999.
- [8] Q. Cai and J. Aggarwal. Tracking human motion in structured environments using a distributed-camera system. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(11):1241–1247, Nov 1999.
- [9] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb. Plan-view trajectory estimation with dense stereo background models. *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, 2:628–635, 2001.
- [10] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2000.*, 2:126–133 vol.2, 2000.
- [11] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):267–282, Feb. 2008.
- [12] M. Harville and D. Li. Fast, integrated person tracking and activity recognition with plan-view templates from a single stereo camera. *International Conference on Computer Vision and Pattern Recognition (CVPR2004)*, 02:398–405, 2004.
- [13] L. Iocchi and R. Bolles. Integrating plan-view tracking and color-based person models for multiple people tracking. *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, 3:III–872–5, 11–14 Sept. 2005.
- [14] M. E. Leventon and W. T. Freeman. Bayesian estimation of 3-d human motion. Technical report, Mitsubishi Electric Research Laboratories (MERL), July 1998.
- [15] H. Sidenbladh, M. J. Black, and D. J. Fleet. Stochastic tracking of 3d human figures using 2d image motion. *Computer Vision, ECCV2000*, pages 702–718, 2000.