A Multi-View Annotation Tool for People Detection Evaluation

Ákos Utasi and Csaba Benedek
Distributed Events Analysis Research Laboratory
Computer and Automation Research Institute, Hungarian Academy of Sciences
Kende u. 13-17, H-1111 Budapest, Hungary
{utasi,bcsaba}@sztaki.hu

ABSTRACT

In this paper we introduce a novel multi-view annotation tool for generating 3D ground truth data of the real location of people in the scene. The proposed tool allows the user to accurately select the ground occupancy of people by aligning an oriented rectangle on the ground plane. In addition, the height of the people can also be adjusted. In order to achieve precise ground truth data the user is aided by the video frames of multiple synchronized and calibrated cameras. Finally, the 3D annotation data can be easily converted to 2D image positions using the available calibration matrices. One key advantage of the proposed technique is that different methods can be compared against each other, whether they estimate the real world ground position of people or the 2D position on the camera images. Therefore, we defined two different error metrics, which quantitatively evaluate the estimated positions. We used the proposed tool to annotate two publicly available datasets, and evaluated the metrics on two state of the art algorithms.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis

General Terms

Design, Verification, Experimentation

Keywords

Multi-view annotation, people detection

1. INTRODUCTION

In many surveillance systems key functionalities involve pedestrian detection and localization in the scene. The location information is used in higher level modules, such as tracking, people counting, restricted zone monitoring, or behaviour analysis. In recent years multi-view surveillance has undergone a great advance, and novel methods have been proposed to improve the efficiency of people detection and localization. However, most of existing multi-view image sequences are annotated using the conventional method of generating 2D bounding boxes around the pedestrians in the images.

This work presents a novel approach for manual ground truth generation for multi-view image sequences and goes beyond the traditional bounding box annotation technique. The proposed tool assumes that the multiple sequences are synchronized and the cameras are calibrated. Moreover, an area of interest (AOI) is also defined by the user on the ground plane. In the proposed annotation the real 3D ground position, the occupancy area on the ground plane (represented by oriented rectangles), and the height is stored for each person.

The rest of the paper is organized as follows. Sec. 2 gives a brief overview of existing annotation datasets for people surveillance. In Sec. 3 the proposed multi-view annotation tool is presented. Moreover, in this section we briefly present two public datasets we manually annotated with our tool. In Sec. 4 we present two different error metrics which can be used to evaluate pedestrian localization algorithms using our annotation data. Our experimental results of two recent methods are presented in Sec. 5. Finally, Sec. 6 concludes the paper.

2. RELATED WORK

Several interactive methods have recently been proposed for multi-view video annotation, however, instead of accurate localization, the main purpose is usually either moving target selection and manipulation [7], or adding pixelwise metadata to the object such as depth estimates or information on material type [11].

Since many recent research works in the literature deal with multi-view people detection, a key issue is to construct a relevant framework to quantitatively validate and compare the different solutions. Although most research articles provide numerical evaluation results, the applied evaluation metrics and the accuracy of ground truth generation face limitations.

Some of the methods calculate the detection errors purely in the image space. The Multiple Plane Model [5] does not exploit metric calibration data for the test sequences; therefore, the authors calculate the distance in pixels between the top of the tracked localization (centroid of top patch of

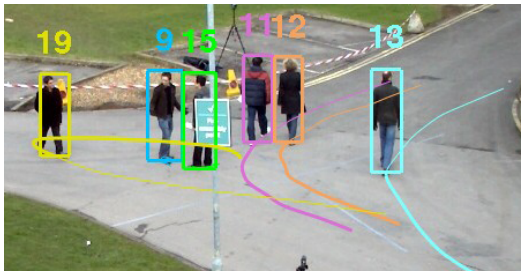

Figure 1: 2D bounding box annotation of the video frames with marking the trajectories of the people, prepared by [1].

the track bounding cubes) and the manually marked top of the heads of the people. The Multiview Sampler Approach [4] optimizes the object configuration in the real 3D space, then it projects the bounding cylinders back to the image planes, creating a bounding box for each target in each view. In the evaluation phase, localization quality is measured by object level precision and recall rates, where a detection is counted as correct if the overlap ratio between the annotated box and the detection box is greater than a threshold. The main problem of the image plane based evaluation techniques is that they provide very limited information about the real accuracy of the model (*e.g.* position error of a localization system), since distances and overlapping areas measured in pixels depend on the distance of the target from the viewpoint and the extrinsic and intrinsic camera parameters. [1] used the CLEAR METRICS for to evaluate their proposed multi-object tracking algorithm, where the tracking precision is calculated by measuring overlap of bounding boxes (see also Fig. 1).

The authors of the Probabilistic Occupancy Map (POM) method [3] follow a different approach: they measure and evaluate the ground positions of the persons by stretching a discrete rectangular grid to the ground plane with typically 25cm resolution, and they attempt to assign to each person a single cell containing his/her ground position. This position error is measured in the real 3D world coordinate system. However, the cell size gives a natural limitation of the accuracy of location estimation. On the other hand, since the target position is represented by a fixed sized cell, we cannot describe different gait phases and body poses with the annotation.

The evaluation methodologies corresponding to the PETS 2009 challenges [2] have included both image space and ground position based quality measures. The organizes considered ground truth annotation with simultaneously defining bounding boxes in all views corresponding to a person, and by locating its 3D position on a discrete grid of 33cm resolution, following the approach of [3]. Since the solutions submitted to this competition were tested on the same input videos, the evaluation rates proved to be relevant for method comparison, however they have still showed the previously mentioned limitations.

For the above reasons, we have proposed a novel validation tool to evaluate various multi-view people localization techniques [3, 14], which describe the location of the people by arbitrary oriented rectangles on the ground planes. With this approach, we can model that people may occupy differently sized and oriented regions in the ground plane, which enable more accurate and comprehensive detection and behavior analysis.

3. PROPOSED ANNOTATION TOOL

From the previous section we can see that in most datasets the annotation is a simple rectangle around the person in the image, but the real 3D position, from a few exceptions, is usually neglected. Moreover, in a conventional annotation process, usually a single view of the scene is displayed, and the user has to manually select the ground truth annotation on that single view. It is obvious, that single view annotation will result in a reduced accuracy when a severe crowd is present in the scene, and the pedestrians become fully or partially occluded.

In order to cope with the above problems we defined a novel annotation format for multi-view sequences, where each pedestrian is annotated with its a) ground rectangle to describe occupancy, and b) height of the pedestrian. However, considering the synchronization and calibration errors of multi-view sequences it is not unambiguous, where the real position of a person is located on the ground. Therefore, we developed a novel tool which displays all the available sequences, the ground rectangle around each person is projected to each camera view, and the user has to manually set the attributes of the rectangle (center position, size and orientation) so that it will contain both legs, while having minimal extent. Thus in our annotation the ground occupancy of each person is represented by a rectangle on the ground plane covering the area of the human body between the two leg positions. The location and the extent of the rectangles can be set up to 1cm accuracy in the world coordinate system. On the other hand, by using multiple camera views the number of missed persons and inaccurately positioned ground truth data caused by occlusions can be reduced significantly, *e.g.* the real position of the person in white shirt in Fig. 2 is hardly visible in the first camera view due to occlusions, but in the second and the third views this person is completely visible.

The graphical interface of the application is simple and provides a straightforward way for ground truth annotation. Fig. 2 demonstrates the key elements of the application's interface. The upper buttons control the actual position in the sequences. The images in the next row display an overview of the whole scene. Here, the AOI is represented by a white rectangle on the ground plane, and the annotation of each pedestrian is visualized using different colors. A new annotation can be created by simply clicking on one of these camera images, in this case the attributes of the annotation are initialized by default values, and the annotation is projected on all views. The application supports the fine tuning of the annotation by displaying the currently selected person in the zoomed images in the next row. Finally, the pane in the bottom contains the control elements for changing the attributes of the ground truth annotation of the selected person. The spin buttons on the left control the x, y ground position of the person and its z height, which corresponds to a line between the hypothetical ground and head positions of the person, and it is perpendicular to the

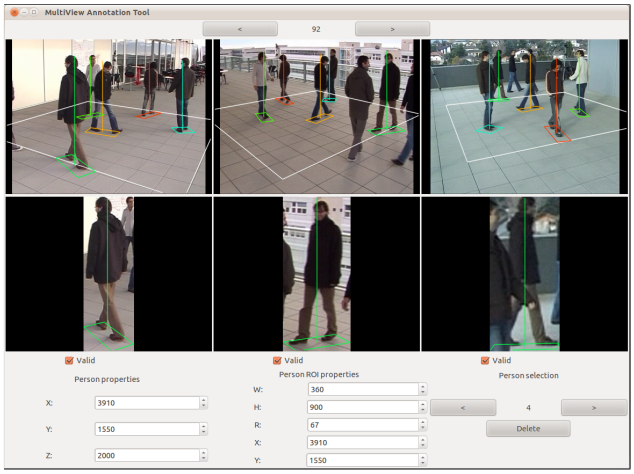

Figure 2: User interface of the Multi-View Annotation Tool. The upper buttons control the position in the image sequence, the upper images show the views of the scene, the bottom images show the selected person, and the controls on the bottom are used for changing the parameters of the ground truth. Ground rectangles and heights are visualized by the projections to all camera views.

ground plane. The spin buttons in the center can be used to change the attributes of the ground rectangle (x, y center position, w, h extent, and r orientation). Finally, the controls on the right can be used for navigating between annotated pedestrians and for deleting annotations from the dataset. Optionally the user can also enable a copy functionality, which is executed upon forward navigation in the image sequence, and it copies all the annotations to the next frame in the sequence, if it does not contain any ground truth data. We found this functionality quite useful for sequences having high FPS rate, where the displacement of the pedestrians between two consecutive frames is small. In this case the user has to manually align the position and the size of the bounding rectangle, which significantly reduces the time required for annotation.

In order to have a cross-platform MVC application we chose the GTK+ widget toolkit to create the graphical interface of our annotation tool, OpenCV library [8] for the image processing tasks, and the calibration software of [12] for projecting the real world coordinates to the images and vica versa. Ground truth data are stored in simple ASCII text files. The application[1] has been tested under Ubuntu Linux 64bit and Windows XP 32bit operating systems.

3.1 Annotated Datasets

We used our tool to create ground truth data for two public multi-view sequences, having different characteristics. First, from the PETS 2009 dataset [9] we selected the *City center* images, which contain approximately 1 minute of recordings (400 frames total) in an outdoor environment. From the available views we selected cameras with large fields of view and we used an AOI of size 12.2m × 14.9m, which is visible from all three cameras, and we annotated all available frames

[1]Available at `http://web.eee.sztaki.hu/~ucu/mvatool`

Figure 3: Example frames from the PETS *City center* sequence. The scene is outdoor and contains moving vegetation and other occluding objects.

in the sequence. Fig. 3 shows sample frames from two of the available three camera views.

The second dataset we used in our experiments is the EPFL *Terrace* [3] dataset (see Fig. 2), which is 3 minutes and 20 seconds long (5000 frames total). This sequence has been annotated in 1Hz frequency resulting in 200 annotated frames. The scene is semi-outdoor, since it was recorded in a controlled outdoor environment and it also lacks some important properties of a typical outdoor scene (*e.g.* no background motion caused by the moving vegetation is present, and no static background objects occlude some parts of the scene). We selected three cameras having small fields of view, and defined the AOI as a 5.3m × 5.0m rectangle.

The two datasets do not only differ in the environment of their scene, but they also have different characteristics with respect to the density of people inside the AOI, as shown in Fig 5. It can be clearly seen that the *Terrace* sequence contains a more severe crowd, and thereby a higher occlusion rate.

During the annotation process we increased the extent of the ground rectangles when the projections had significant difference in the three selected camera views (caused *e.g.* by synchronization error or calibration inaccuracy, see Fig. 4). Moreover, it is difficult to decide visually whether the person near the borders of the AOI should be considered as being inside or outside. Therefore, we created our annotations in a larger AOI using an additional 25cm buffer zone, which decreases the false detections near the border area.

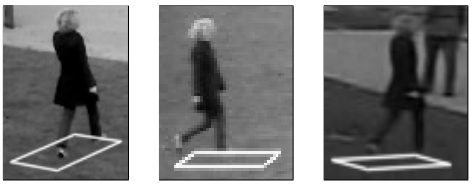

Figure 4: Sample frames from the *City center* sequence show significant synchronization error and calibration inaccuracy.

4. EVALUATION METHODOLOGY

As discussed in Sec. 3 our annotations at a given timestep are the ground occupancies of the pedestrians represented by an $\mathbf{R} = \{r_1, \ldots, r_m\}$ set of m rectangles on the ground plane, where each r_i rectangle is parametrized by its ground position $x(r_i), y(r_i)$, size $w(r_i), h(r_i)$ (width and height), and

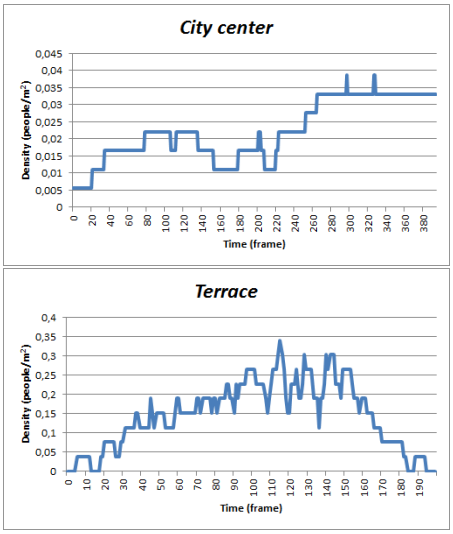

Figure 5: Comparison of the dynamics of the two datasets: people density over time in the *City center* (top), and in the *Terrace* (bottom) sequences. A more severe crowd and a higher occlusion rate are present in the *Terrace* sequence.

orientation $\theta(r_i)$. However, different methods estimate the position of the pedestrians using different models. Moreover, some methods estimate the real world location, while other existing methods estimate the location in the camera image. To cope with the first problem the proposed evaluation methodology requires the evaluated method to present the estimated location of a p person as an $x(p), y(p)$ ground coordinate (in real world or in the image). This can be performed in a straightforward way in any method. For the second problem we defined two different error metrics, which will be discussed later in detail.

Let us assume that the set of detected people at a given timestep is denoted by $\mathbf{P} = \{p_1, \ldots, p_n\}$. Given the ground truth data $\mathbf{R}$ and the estimated positions $\mathbf{P}$ we define a match function $m(i,j)$ to indicate whether the estimated p_j is inside the ground truth annotation r_i or not, *i.e.*

$$m(i,j) = \begin{cases} 1 & \text{if } p_j \text{ is inside } r_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and we use the Hungarian algorithm [6] to find the maximum matching, *i.e.* the maximum utilization of $\mathbf{M} = [m(i,j)]_{m \times n}$. We denote by $\mathbf{A} = [a(i,j)]_{m \times n}$ the assignment obtained by the algorithm, *i.e.*

$$a(i,j) = \begin{cases} 1 & \text{if } p_j \text{ was assigned to } r_i \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Finally, we count

- Missed Detections:

$$\mathrm{MD} = \#\left\{ r_j : \sum_{i=1}^{n} a(i,j) = 0 \right\} ,$$

i.e. no estimation was assigned to the ground truth (represented by white rectangles with black outline in Fig. 6);

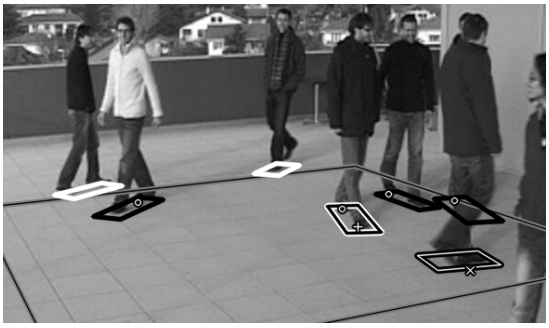

Figure 6: Rectangles represent the ground truth, the ∘, +, and × symbols denote the position estimates. Bold black rectangles with a ∘ symbol denote the successful detections, white rectangles with black outline represent missed detections, black rectangles with white outline and a + symbol show the multiple instances, and the × symbols denote the false detections. Rectangles partially inside the AOI are denoted by bold white rectangles. The AOI is represented by a black rectangle with gray outline.

- False Detections:

$$\mathrm{FD} = \#\left\{ p_i : \sum_{j=1}^{m} a(i,j) = 0 \right\} ,$$

i.e. no ground truth could be assigned to an estimate (represented by × symbols in Fig. 6);

- Multiple Instances:

$$\mathrm{MI} = \sum_{j=1}^{m} \max\left(0, \sum_{i=1}^{n} a(i,j) - 1\right) ,$$

i.e. multiple estimates were assigned to a ground truth (represented by black rectangles with white outline and a + symbol in Fig. 6);

- Total Error:

$$\mathrm{TE} = \mathrm{MD} + \mathrm{FD} + \mathrm{MI} .$$

Finally, we neglected the MDs if the ratio of the area of r_j inside the AOI and the total area of r_j does not exceed 50% (*i.e.* when a person is near the borders of the AOI, it is difficult to decide if he is inside or outside, this is represented by a bold white rectangle in Fig. 6). Note that this step reduces the MDs occurring near the borders of the AOI.

4.1 Error Metrics

The three error types defined above can be computed either from the real word or from the image coordinates. Therefore, we defined two different comparison metrics by determining $\mathbf{M}$ and $\mathbf{A}$ from

1. the real world ground truth and position estimates: we call this the Ground Position Error (GPE) metric;
2. the projected ground truth and positions, with selecting the view with the minimal TE: called the Projected Position Error (PPE) metric.

These two tests allow other methods to be compared against each other whether they estimate the real world ground coordinate of people (*e.g.* [3, 14]) or the 2D position on the camera images (*e.g.* using camera homography instead of calibration, such as [5]).

Finally, after counting all the false localization results (MD, FD, MI) on all annotated frames we express them in percent of the number of all objects, we denote these ratios by MDR, FDR, MIR, and TER. Note that while MDR ≤ 1 and MIR ≤ 1 always hold, in case of many false alarms FDR (thus also TER) may exceed 1.

5. USE CASES

After defining the error metrics we numerically evaluated two methods using our ground truth data. The input of both methods are binary images corresponding to the results of background-subtraction from multiple camera views. In our experiments we used the adaptive mixture of Gaussians approach of [10], also used by the authors of [14].

The first approach is the 3D Marked Point Process (3DMPP) model [14]. This method in the first step extracts pixel-level features from the projected foreground masks to estimate the leg positions of the pedestrians on the ground plane. This information with additional prior geometrical constraints is embedded into a 3D object configuration model, where each person is modelled by a cylinder object. The final configuration results are obtained by an iterative stochastic energy optimization algorithm. The parameters of this technique include $\hat{f}$, which controls dynamic range of the pixel-level feature, and d_0 the minimal feature value required for object acceptance (the notation $D = 1/d_0$ is also used). Moreover, in our experiments the R radius of the cylinder objects were set to a fixed 40cm value. More details can be found in [13, 14].

The second method is the Probabilistic Occupancy Map (POM) technique [3], which is a generative method and divides the ground plane of the AOI into a discrete 2D grid, having a predefined resolution (typically in the $10-40$cm range). The method estimates the marginal probabilities of presence of pedestrians at every grid location under a simple appearance model, which is parametrized by a family of rectangles approximating the silhouettes of average sized individuals (with height 175cm and width 50cm) standing at every grid position, from every camera view. The POM method outputs grid position occupancy probabilities, and in our experiments people position estimates are obtained by thresholding this probability map. Thus the parameters of this method are the ν grid resolution, and the τ threshold value. More information can be found in [3].

According to our experiments multiple detections are in general the least frequent artefacts (*i.e.* MIR $\ll$ TER), therefore we evaluated the effects of the different parameter settings on the MDR and on the FDR only. The resulting values are presented in Fig. 7 for both the *City center* (top) and the *Terrace* sequences (bottom), and they can be used for fine tuning the methods by selecting the optimal parameter values, which minimize *e.g.* the TER.

To demonstrate the strong connections between the GPE and the PPE metrics, we have displayed in Fig. 8 the MDR and FDR plots obtained by the both metrics with the same parameter settings using the 3DMPP method for the *City center* sequence. The similarity of the corresponding plots confirm that the two metrics are equivalently appropriate for method evaluation, thus PPE can be used for techniques where camera calibration information is not available.

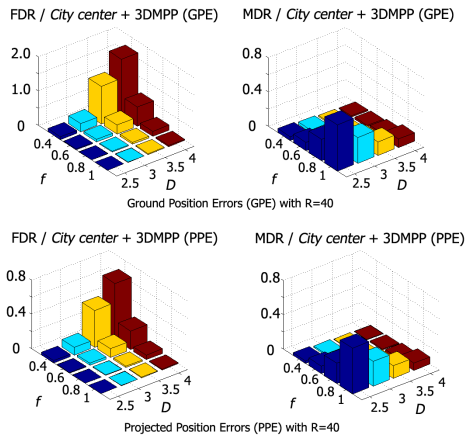

Figure 8: Comparison of the GPE and the PPE metrics using the 3DMPP model for the *City center* sequence. FDR and MDR plots of both metrics are shown using various parameter values. Similarity of the corresponding plots confirm the appropriateness of both GPE and PPE for method comparison.

6. CONCLUSIONS

In this paper we presented a novel annotation format for evaluation multi-view people detection methods. Instead of using a conventional bounding box annotation, our ground truth data represents the location of a pedestrian by a rectangle on the ground plane of the scene. To create such annotations we developed a multi-view annotation tool, which helps the user in the annotation process of crowded scenes by displaying the multiple camera views of the scene. We provided ground truth annotation for two public multi-view sequences having different characteristics. Then we defined two different metrics for the evaluation of people detection algorithms on our datasets. Finally, we selected two methods for multi-view people detection and performed experiments using our ground truth data. A future extension of both the annotation format and the application might involve the ability to represent the correspondences between pedestrians in the consecutive video frames, and thereby providing an extended ground truth dataset for evaluating people tracking algorithms.

7. ACKNOWLEDGEMENTS

This work was partially supported by the PROACTIVE (PRedictive reasOning and multi-source fusion empowering AntiCipation of attacks and Terrorist actions In Urban EnVironmEnts) project of the EU. The work of the second author was supported by the Hungarian Research Fund (OTKA #101598), and by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

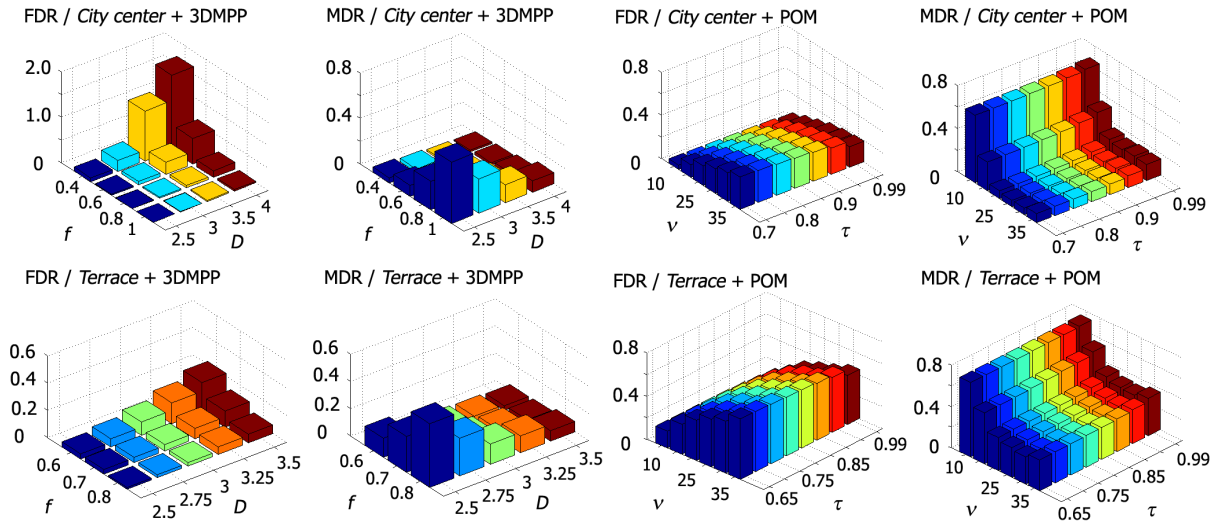

Figure 7: Evaluation of the 3DMPP and the POM method for the *City center* (top) and the *Terrace* (bottom) sequences using the GPE metrics. Both the FDR and the MDR plots are shown for various parameter settings.

8. REFERENCES

[1] A. Andriyenko and K. Schindler. Globally optimal multi-target tracking on a hexagonal lattice. In *Proceedings of the 11th European Conference on Computer Vision*, pages 466–479, Heraklion, Greece, Sept. 2010.

[2] A. Ellis, A. Shahrokni, and J. M. Ferryman. PETS 2009 and Winter-PETS 2009 results: A combined evaluation. In *Proceedings of the 12th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–8, Snowbird, UT, USA, Dec. 2009.

[3] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):267–282, 2008.

[4] W. Ge and R. T. Collins. Crowd detection with a multiview sampler. In *Proceedings of the 11th European Conference on Computer Vision*, pages 324–337, Heraklion, Greece, Sept. 2010.

[5] S. M. Khan and M. Shah. Tracking multiple occluding people by localizing on multiple scene planes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(3):505–519, 2009.

[6] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[7] G. Miller, S. Fels, A. Al Hajri, M. Ilich, Z. Foley-Fisher, M. Fernandez, and D. Jang. MediaDiver: Viewing and annotating multi-view video. In *Proceedings of the 30th Conference on Human Factors in Computing Systems*, pages 1141–1146, New York, NY, USA, May 2011.

[8] OpenCV. Open Source Computer Vision Library. http://opencv.willowgarage.com/.

[9] PETS. Dataset - Performance Evaluation of Tracking and Surveillance, 2009. http://www.cvg.rdg.ac.uk/PETS2009/a.html.

[10] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[11] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, and L. Van Gool. Using multi-view recognition and meta-data annotation to guide a robot's attention. *International Journal of Robotics Research*, 28(8):976–998, 2009.

[12] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, 1987.

[13] Á. Utasi and Cs. Benedek. Multi-camera people localization and height estimation using multiple birth-and-death dynamics. In *Proceedings of The 10th International Workshop on Visual Surveillance*, pages 74–83, Queenstown, New Zealand, Nov. 2010.

[14] Á. Utasi and Cs. Benedek. A 3-D marked point process model for multi-view people detection. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, pages 3385–3392, Colorado Springs, CO, USA, June 2011.