# A STATISTICAL METHOD FOR OBJECT LOCALIZATION IN MULTI-CAMERA TRACKING

*László Havasi, Zoltán Szlávik*

Computer and Automation Research Institute
H-1111, Kende u. 13-17, Budapest, Hungary
{havasi, szlavik}@sztaki.hu

## ABSTRACT

The aim of the paper is to show that localization of observed moving objects is possible in a multicamera environment, based on the simple foreground mask and the estimation of the size distribution on image plane. The method is flexible, it can handle arbitrary number of cameras and objects. It is based only on change detection masks and does not depend on appearance information. Due to its statistical nature the proposed method efficiently handles such challenging situations as changes in viewpoint, occlusions due to view variations, background clutter.

***Index Terms***— localization, surveillance

## 1. INTRODUCTION

The estimation of location and inter-camera correspondence of multiple objects in a surveillance system is closely related to the task of tracking multiple targets. Localization and tracking of multiple targets in crowded environments is a challenging task due to inter-object occlusion between targets. Single-object single view detection and tracking is much simpler, the perceived visual attributes of the target, like appearance, shape, and motion are not changing much during continuous observation. Having more targets increases inter-object occlusions, a detected blob in the image may belong to several targets in the real 3D space. There will be a number of situations when the tracking will be impossible, due to the inter-object occlusions, using only a single view. A natural way to overcome inherent problems of single view tracking is the use of multiple cameras with overlapping fields of view for the observation of a dynamic environment. By using multiple cameras the number of possible inter-object occlusions is reduced and a more precise tracking and estimation of objects locations is possible.

The present paper deals with the problem of localization of multiple targets in multiple camera surveillance systems. The localization of the objects inherently solves the problem of inter-camera correspondence. The proposed method uses only the output of a background-foreground segmentation algorithm as input and it is not based on any sophisticated appearance, colour or shape model of the observed objects.

## 2. RELATED WORK

In [6] a tracking algorithm was presented to track multiple objects in multiple views by modelling the appearance of blobs using colour histogram techniques. In [1] a Bayesian network was used to combine epipolar geometry and appearance based modalities to match multiple objects in multiple views. Bayesian networks were also used in [2], to track objects and resolve occlusions across multiple calibrated cameras. [5] uses stereo cameras and fuse the information from multiple views in 3D space. These and similar methods are heavily based on the quality of features detected (appearance, shapes etc.) and they will fail if the image primitives could not be reliably detected.

Depth-based object-ground segmentation and object localization techniques are based on the idea that a human usually stands out in a 3D environment. [14] presents a method for human detection by comparing the appearance information in each camera view and combining them based on a region based stereo method. [15] uses silhouettes from different views in order to interpret the visual hull. The incorrect detections are filtered using size and temporal criterions. [13] also uses silhouette information to detect feet positions in camera images and then maps detected positions to a virtual ground image.

A recent idea of homographic occupancy constraint presented in [4] and [8] fuses information from multiple views using geometrical constructs and resolves occlusions by localizing people on multiple scene planes. The presented method is image based, it uses detected blobs and requires only 2D constructs like planar homographies to perform fusion in the image plane without requiring to go in 3D space.

## 3. OUR APPROACH

Our approach has similar aspects to that of presented in [8]. There is no direct correspondence estimation between different views. The basic assumptions are:

- the observed objects are moving on a groundplane;

- the distribution of the imaged size of moving targets is given on image plane [9];
- the camera model is perspective;
- the homography between the views is known;
- object position is defined with the location where feet touche the groundplane.

The overall aim of the proposed method is to estimate for each pixel in the image plane the probability of being an image of groundplane location occupied by an observed object (in case of humans this is the probability of observing their feet) - the "occupancy probability" in short. The basic idea of our method is to observe the dynamics of a scene for a while and learn the distribution of the size of the observed moving objects. Based on this and the camera-scene geometry we estimate for every image pixel the occupancy probability.

## 4. LEARNING SIZE DISTRIBUTION OF OBSERVED OBJECTS ON THE IMAGE PLANE

The approximate knowledge of object sizes on the image plane is proven to be useful information for geometrical scene analysis tasks [9]. The learning of the distribution of the imaged size of the moving objects over the image is based on the work [9]. The estimated motion statistics contains the probability of motion co-occurences between image's pixels and the average size, height and width, of the observed objects also can be extracted. Examples of such extraction are shown in Fig. 1. The details of the estimation can be found in [9]. In the followings the average object size (height and width) is defined as $S_I(x)$, where $I$ is the camera index ($I$=1…number of cameras) and the $x$ is the position in the image.

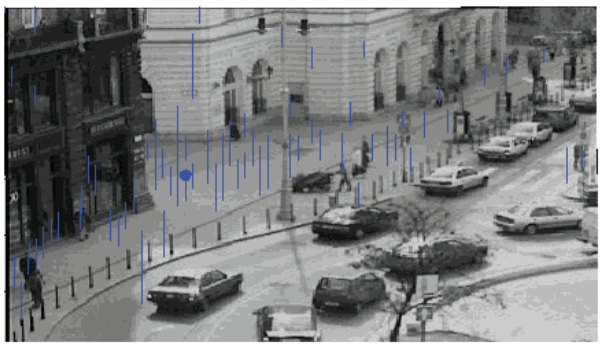

Figure 1. Examples of estimated heights for a real world image sequence [9].

## 5. ESTIMATION OF OCCUPANCY PROBABILITY FOR SINGLE VIEW

Let us suppose that motion is detected at an arbitrary position $x$ in image plane $I$. For motion detection we use a version of the well known Stauffer-Grimson algorithm [10] and a binarized change detection mask is used in the latter computations:

$$m_I(x) = \begin{cases} 1, where\ motion\ is\ detected \\ 0, otherwise \end{cases}$$

Having no information about neighbouring pixels (no object information) and due to the interobject occlusion the detected foreground might belong to multiple objects in world space along the line of sight till the ground plane.

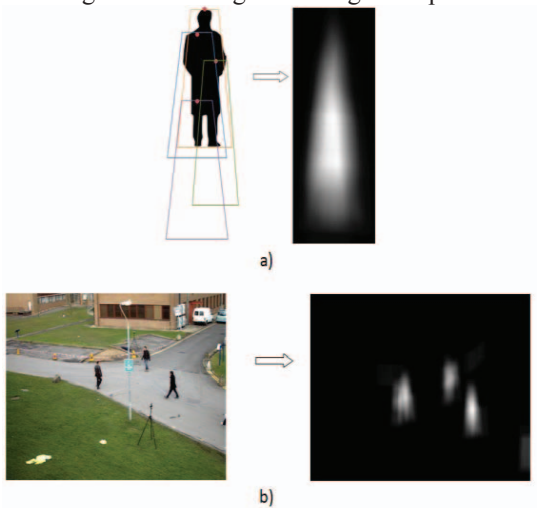


Figure 2. The scheme of estimation for $L_I$ occupancy probabilities is illustrated in *a)*. Example of the estimation of $L_I$ for a real image could be seen in *b)*.

In order to estimate the possible groundplane locations the algorithm makes use of the estimated size information, which is available for all the pixels of the image plane $I$. The detected foreground may refer to the top of an observed object or to the bottom, the location on the ground, or any part between. Let us assume that the corresponding average sizes of the observed objects at the two extreme cases are $(h_1,w_1)$ where the foreground corresponds to the location on the ground and $(h_2,w_2)$ where the foreground corresponds to the top of the observed object. This means that the trapezoid which is the image of the possible groundplane locations $L$, has dimensions: height $h_2$, base length $w_1$ and $w_2$. Let $I(L_x)$ denotes the set of image pixels that are the images of possible groundplane locations $L_x$ related to a given $x$.

We define an image sized binary mask for every image pixel $\vec{x}$ with $T_{x,S_I}$ wich has a value in pixel $y$:

$$T_{x,S_I}(y) = \begin{cases} 1, y \in I(L_x), \\ 0, y \notin I(L_x). \end{cases}$$

The summarization of such masks on a given camera frame $I$ leads to a probability map that defines possible object positions on the image for a given motion mask:

$$L_I = \sum_x T_{x,S_I}\, m_I(x)$$

Figure 3. Example of the estimated $L_\Sigma$ occupancy locations. Some of the objects and their corresponding locations are indicated.

Figure 2 demonstrates these computation steps.

The estimated $L_I$ depends on the imaged size of the moving objects. This means that locations of objects that are closer to the imaging device get more votes than those being farther due to the perspective camera model. In order to make the estimation size invariant a size dependent scaling factor is introduced. Let us assume that in an ideal case (when the moving object is ideally separated from the background) the value of the map should be 1. The scaling factor for the pixel $x$, where the corresponding $I(L)$ is defined by the triple $(h_1, w_1, w_2)$, calculated as $1/( h_1 w_1)$. By using this height dependent scaling factor near and distant objects' locations will have equal weight in the fusion space $L_\Sigma$. The result is invariant to size variations that are caused by camera-object distance variations and the perspective camera model.

## 6. LOCALIZATION OF OBJECTS IN MULTIPLE VIEWS

Let us suppose a scene with multiple cameras with overlapping fields of view and that the groundplane induced homographies between cameras are known or the homographies between cameras and the groundplane are known. Warping the images of different views onto each other with such homographies the images of occupancy locations of the same object will consistently warp to each other. Thus we easily can select a common frame for all the cameras and fuse the estimations on this common frame. For easier understanding and easier presentation let us suppose that we have the latter case. In such case for a given $L_I$ the corresponding ground location is determined by $\bar{L}_I = H_I^{-1} L_I$, where $H$ is a homography between image $I$ and the groundplane - the common frame -, $G$. Having estimated $L_I$ for all the cameras (images $I$), the corresponding $L_\Sigma$ on the common frame is the aggregation of all the warped $L_I$ made for the common frame locations in different images $I$:

$$L_\Sigma = \sum_I H_I^{-1} L_I$$

$L_\Sigma$ is normalized to have maximum value of $1$. Example of an estimated $L_\Sigma$ is shown in Fig. 3.

Where, higher value means higher probability of a true object location. Note that local maxima correspond to object locations in the different views. For easier understanding several objects in the different views and their corresponding common frame locations are shown.

$L_\Sigma$ will have local maxima at positions of the most possible locations of the observed objects. Due to using a size dependent scaling factor in the estimation of $L_I$ in an ideal case, when the moving objects are almost perfectly separated from the background, the local maxima that correspond to the locations of observed objects should be close to *1*. By detecting its local maxima locations of the observed objects can be recovered.

The value *T=0.4* means that all the object locations are detected where at least 40% of the object mask is detected. By subtracting $L_I$ that correspond to the actual global maximum, the subtracted $L_I$ has no more effect on $L$. This means that possible false positive locations are also removed and the algorithm is more robust to these errors.

**Object localization algorithm**

1. Initialize object imaged height $S_I(x)$ maps.
2. Occupancy probability estimation:
   - Determine foreground masks $m_I(x)$ for all views.
   - Estimate imaged occupancy probabilities $L_I$ for all views.
   - Summarize all transformed $L_I$ into $L_\Sigma$
3. Localization steps:
   - Detect global maxima of $L_\Sigma$
   - Subtract $L_I$ parts that corresponds to the detected global maxima.
   - Repeat until the value of the global maxima reaches *T=0.4*

## 7. EXPERIMENTAL RESULTS

The proposed approach for localization has been tested on PETS 2009 S2 dataset. Example results for different processing steps of the method can be seen in Figures 2., 3. In Figure 3. some of the estimated location statistics are shown. It can be seen that local maxima of statistics are corresponding to different objects in 3D space and backprojecting to the image plane gives the correspondence of the detected objects across different views. The example clearly shows that the model is able to handle such a challenging situation as the inter-object occlusion in a single view.

## 8. CONCLUSIONS

We have described an approach for the localization of moving objects in a multi camera surveillance system, which allows the localization the observed moving object without appearance investigation. Based on a simple scene geometry model and the learned size distribution on image plane of observed moving objects an effective view fusion has been given for the improvement of multiple object detection. The fusion of several views helps to remove ghosts and false detections and to resolve inter-object occlusions.

## 9. ACKNOWLEDGEMENT

The authors would like to acknowledge the support received from the Hungarian Research Fund under the contract PD76414 and from the Bolyai Postdoc Research Scholarship.

## 10. REFERENCES

[1] T. Chang and S. Gong, “Tracking Multiple People with a Multi- Camera System,” Proc. IEEE Workshop Multi-Object Tracking, 2001.

[2] S. Dockstader and A. Tekalp, “Multiple Camera Fusion for Multi- Object Tracking,” Proc. IEEE Workshop Multi-Object Tracking, 2001.

[3] M. Han, W. Xu, H. Tao, and Y. Gong, “An Algorithm for Multiple Object Trajectory Tracking,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004.

[4] S.M. Khan and M. Shah, “A Multi-View Approach to Tracking People in Crowded Scenes Using a Planar Homography Constraint,” Proc. Ninth European Conf. Computer Vision, 2006.

[5] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, “Multi-Camera Multi-Person Tracking for Easy Living,” Proc. Third IEEE Int’l Workshop Visual Surveillance, 2000.

[6] J. Orwell, S. Massey, P. Remagnino, D. Greenhill, and G. Jones, “A Multi-Agent Framework for Visual Surveillance,” Proc. IEEE Int’l Conf. Image Processing, 1999.

[7] A. Yilmaz, O. Javed, and M. Shah, “Object Tracking: A Survey,” ACM J. Computing Surveys, 2006.

[8] M. S. Khan, M. Shah, "Tracking Multiple Occluding People by Localizing on Multiple Scene Planes", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 31(3), pp. 505 - 519, 2009.

[9] L. Havasi, T. Szirányi, "Extraction of horizontal vanishing line using shapes and statistical error propagation", Symposium of ISPRS Commission III. Photogrammetric computer vision, pp. 167-173, 2006.

[10] C. Stauffer, E. Grimson, “Learning Patterns of Activity Using Real-Time Tracking”, *IEEE Trans. PAMI*, vol. 22(8), pp. 747-757, 2000.

[11] T. B. Moeslund, A. Hilton, V. Krüger, "A survey of advances in vision-based human motion capture and analysis", *Computer Vision and Image Understanding*, vol. 104(2), pp. 90-126, 2006.

[12] Y.A. Ivanov, A.F. Bobick, J. Liu, "Fast lighting independent background subtraction", *International Journal of Computer Vision*, vol. 37(2), pp. 199–207, 2000.

[13] S. Iwase, H. Saito, "Parallel tracking of all soccer players by integrating detected positions in multiple view images", Proc. ICPR, pp. 751-754, 2004.

[14] A. Mittal, L.S. Davis, "M2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene using region based stereo", Proc. ECCV, 18-33,2002.

[15] D.B. Yang, H.H.G. Banos, L.J. Guibas, "Counting people in crowds with a real-time network of simple image sensors", Proc. ICCV, 122-129, 2003.