LEARNING THE DEPTH STRUCTURE OF A MONITORED SCENE

J. Renno, D. Greenhill, S. Velastin, G.A. Jones

Digital Imaging Research Centre, Kingston University
Penrhyn Road, Kingston upon Thames, Surrey, UK KT1 2EE
{j.r.renno,d.greenhill,sergio.velastin,g.jones}@kingston.ac.uk
www.kingston.ac.uk/dirc

ABSTRACT

Complex scenes such as underground stations and malls are composed of static occlusion structures such as walls, entrances, columns, turnstiles, *etc*. In any object tracking application, these structures induce *static occlusions* which can significantly degrade the performance of tracking and the subsequent interpretation of the monitored scene. Handling such occlusions relies on predicting when such occlusions are likely to occur *i.e.* requires an understanding of the 3D structure of the scene. A method for automatically learning the depth structure of the scene is described which utilises observations from the monitored scene itself.

1. INTRODUCTION

Within tracking, *occlusion analysis* refers to two largely distinct processes. First, the *dynamic* occlusion of one moving object by other moving objects which causes particular difficulties in continuing to establish the temporal identity. Second, *static occlusion* involves the interaction between moving objects and static occluding structures with in the scene such as walls, entrances, columns, turnstiles, barriers *etc* or the image boundary itself. These structures induce *partial occlusion* which can significantly impair the accuracy of the data association process, or *full occlusion* for different lengths of time during which the object completely disappears. Both static and dynamic occlusion processes are handled elegantly by using depth to reason about occlusion. As an object proceeds through the scene, inter-frame correspondence - which typically involves a spatio-chromatic comparison of observation pixels with an *appearance model* - becomes problematic where there has been significant occlusion of the observation. The depth map allows the visibility of the object to be predicted and, by incorporating this visibility into the comparison, improves the accuracy of correspondence. Recently Senior[6] automatically estimates an *occlusion map* defined as pixels modelled as part of the reference image but which were never occluded. Of course, occluding surfaces can be also be occluded by moving objects. Using bounding boxes, Ellis and Xu manually identified *long-term*, *short-term* and *border* occlusions, and employed a Bayesian network to infer the status of unmatched blobs as they interacted with these occlusion structures[2]. However, the most effective occlusion representation is in fact a depth map. Using a similar philosophy to that proposed in this paper, Schödl and Essa used detected moving objects to infer the relative depth structure[5]. Since these blobs lacked any depth, an extremely time-consuming search process based on minimum description length was used to partition the image into relatively few depth planes.

2. LEARNING THE DEPTH LANDSCAPE

This section describes the establishment of the scene structure by generating the depth probability density functions (PDFs) at each pixel from a training set of detected blobs. The connected region of pixels associated with each observation of a person occludes some static scene element such as a wall of unknown depth. Such an observation constrains the occluded structure to lie at some distance beyond the observation. Thus, assuming that the training set of observed people explore all navigable space, the union of their 3D trajectories will demark the depth structure of the scene. After regularisation, the depth map may be used to support reasoning about static occlusions.

2.1. Defining the relative depth of an observation

The *ground plane* relates the depth of a person to the projected image position of the feet of that person. Thus the row position correlates inversely with depth from the camera. Rather than relying on performing an unnecessary calibration, the row position measured from the bottom of the image is used as a non-linear proxy for the actual depth. Thus the relative depth D of any object in contact with the ground plane at pixel row i is defined as

$$D(i) = N - i \tag{1}$$

where N is the height of the image.

In most cases, imagery generated by cameras mounted high on walls or ceilings contain objects whose head and shoulders are not usually occluded by the scene furniture. For occluded observations, depth estimates need to be recovered from the unoccluded head. The row position i_t of the top of a foreground blob is related to the unobserved location of the feet $\widehat{i}_b$ by the height μ *i.e.*

$$\widehat{i}_b = i_t + \mu \tag{2}$$

The projected pixel height of a person in an image plane is a function of the vertical position of the person in the image. For typical camera installations, this relationship is practically linear *i.e.*

$$\mu = \gamma(i_b - i_h) \tag{3}$$

where i_b is the position of the feet, γ is the expansion rate, and i_h is the *horizon*[4]. Thus assuming that a person's head is unoccluded in typical scenes, it is possible to estimate the occluded location of the feet $\widehat{i}_b$ (and hence depth D) from the head location by combining equations 3 and 2 to generate

$$\widehat{i}_b = \frac{i_t - \gamma i_h}{1 - \gamma} \tag{4}$$

The minimum value for D is defined as the unobservable row position of the feet of an average-sized person whose head is located just below the bottom edge of the image. The most distant object would be located at the horizon i_h. Thus the range $[D_{\min}, D_{\max}]$ of D is defined as

$$D_{\min} = N - \gamma i_h (1-\gamma)^{-1}, \qquad D_{\max} = N - i_h \tag{5}$$

Since many objects are not of average height, it is preferable to estimate the depth D of an object from the base of its foreground blob when unoccluded. We validate the actual blob base i_b by requiring that the difference in real and predicted blob heights $|\widehat{i}_b - i_b|$ is less than some proportion τ of the predicted height. Thus the depth of any detected foreground object is computed as follows

$$D = \begin{cases} N - i_b; & \text{if } |\widehat{i}_b - i_b| < \tau(\widehat{i}_b - i_t), \\ N - \frac{i_t - \gamma i_h}{1-\gamma}; & \text{else.} \end{cases} \tag{6}$$

2.2. Constructing Depth Probability Density Functions

A priori, the depth D of any background pixel ϕ at row i_ϕ in the image plane is assumed to belong to a uniform density function between the limits $[D_{\min}, D(i_\phi)]$ *i.e.* it cannot lie at a distance greater than the ground floor element that projects to that pixel. All pixels lying within the mask of a moving person ω can be assigned a distance D_ω given by equation 6. Such observations of moving people eliminate depths from the pixel's PDF which are closer than the observed object. For each image pixel ϕ we maintain an observed depth histogram $z_\phi(D)$ in the range $[D_{\min}, D(i_\phi)]$ which we increment at the depth associated with any blob which contains the pixel. These histograms are computed from the training set Ω of all moving objects.

$$z_\phi(D) = \sum_{\omega \in \Omega_\phi} \delta(D - D_\omega) \tag{7}$$

where $\delta(\cdot)$ is the unit impulse function, and $\Omega_\phi \subset \Omega$ is the set of blobs containing the pixel ϕ. Integrating each pixel's observation histogram over D generates the activity map A_ϕ

$$A_\phi = \sum_{D=D_{\min}, D(i_\phi)} z_\phi(D) \tag{8}$$

Figure 1(b) illustrates the activity recovered from 15,000 frames of the *Underground* scene in Figure 1(a).

(a) Underground Scene (b) Activity Map

Fig. 1. Underground Scene and Activity Map

From the depth observation histograms, we generate a probability density function $\Pi_\phi(D)$ of the likely depth of any static scene point projected to each pixel by accumulating over D and normalising as follows

$$\Pi_\phi(D) = \frac{1}{A_\phi} \begin{cases} \sum_{d=D_{\min}}^{D} z_\phi(d) & \text{if } D < D(i_\phi), \\ 0 & \text{else.} \end{cases} \tag{9}$$

Figure 2 presents PDFs for a number of locations in the underground scene of Figure 1(a). Typically these PDFs exhibit a plateau beyond the deepest observation. Thus the subtraction of a small linear function is used when building an initial depth field Z_ϕ^0 to bias the result towards the closest viable depth *i.e.* $Z_\phi^0 = \text{argmax}_D \ \{\Pi_\phi(D) - \Delta.D\}$.

2.3. Regularization of Depth

The depth probability density functions for each pixel are noisy as the paths taken by individuals do not necessarily visit all the available ground plane. Fragmentation and merging of detected regions also generates erroneous depth estimates. As a result it is necessary to regularize the depth fields and casting the problem as one of optimal assignment

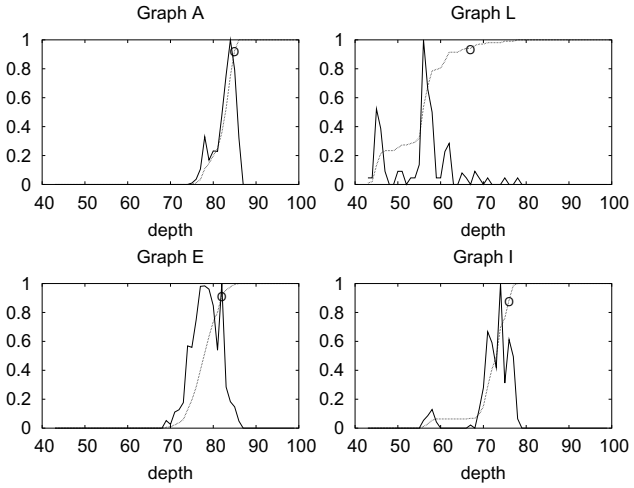

Fig. 2. Probability Density Functions

of depth labels using a Markov Random Field approach[1, 3]. Representing label continuity between neighbouring pixels as a Gibbs probability distribution facilitates the derivation of relatively simple iterative labelling within a Bayesian probability framework.

The selection of depth labels is formulated within a MAP framework in which pixel depth PDF information and label smoothness constraints may be embedded. Let Λ be a particular depth labelling of all pixels $\phi \in \mathcal{I}$ and Ω be the set of depth observations. The maximum *a posteriori* probability rule for selecting the correct image labelling Λ^* may be written in the following form

$$\Lambda^* = \text{argmax}_\Lambda p(\Omega|\Lambda) Pr(\Lambda) \tag{10}$$

The prior probability $Pr(\Lambda)$ of a labelling is modelled by a Gibbs distribution

$$Pr(\Lambda) = \frac{1}{Z} \exp\left(-\frac{1}{T}\varepsilon(\Lambda)\right) \tag{11}$$

where Z and T are the *normalisation* and *temperature* terms. The energy term $\varepsilon(\Lambda)$ is formulated to penalise label configurations which contain regions of non-homogeneous labelling or irregular region boundaries *i.e.* measures the consistency of pixel ϕ having depth λ and its neighbour ϕ' having depth λ'. In general, to encourage locally smooth interpretations, large energy values should be generated for depth discontinuities **unless** these pixels straddle an occluding boundary. Such occluding boundaries can be signalled by discontinuities in the activity map of equation 8 - see Figure 1(b). Thus, given the gradient magnitude E of the activity field A, the energy term is defined as

$$\varepsilon(\Lambda) = \sum_{\phi \in \mathcal{I}} \sum_{\phi' \in \mathcal{N}_\phi} \Delta\varepsilon\left(\lambda_\phi, \lambda'_{\phi'}\right) \tag{12}$$

$$\Delta\varepsilon\left(\lambda_\phi, \lambda'_{\phi'}\right) = \frac{\sigma_E}{\sigma_\lambda} \cdot \frac{|\lambda - \lambda'|}{(\sigma_E + \max(E_\phi, E_{\phi'}))} \tag{13}$$

where $\mathcal{N}_\phi$ are the neighbours of pixel ϕ, and σ_E^2 and σ_λ^2 are the variances of the gradient magnitude E and depth map respectively. Assuming that observations Ω are independent and that the pixel depth PDFs $\Pi_\phi(D)$ are known, equation 10 can be simplified by taking the logarithm

$$\Lambda^* = \text{argmax}_\Lambda \sum_{\phi \in \mathcal{I}} \left\{ \ln \Pi_\phi(\lambda) - \frac{1}{T} \sum_{\phi' \in \mathcal{N}_\phi} \Delta\varepsilon\left(\lambda_\phi, \lambda'_{\phi'}\right) \right\}$$

As it is prohibitively expensive to search over all possible image label configurations, the functional is usually optimised by the iterative application of the following pixel update rule[3].

$$\hat{\lambda}_\phi^{(t)} = \text{argmax}_\lambda \left\{ \ln \Pi_\phi(\lambda) - \frac{1}{T} \sum_{\phi' \in \mathcal{N}_\phi} \Delta\varepsilon\left(\lambda_\phi, \hat{\lambda}_{\phi'}^{(t-1)}\right) \right\}$$

where $\hat{\lambda}_\phi^{(t)}$ is the best new label for pixel ϕ and $\hat{\lambda}_{\phi'}^{(t-1)}$ is best previous label for pixel ϕ'. This iterative procedure is terminated when the labelling stabilises.

Figure 3 presents the regularized depth of the *Underground* scene of Figure 1(a). The *stepped* depth of the ticket barriers in the Underground has become spatially coherent. In both sequences, the ground plane has become much more evident. However both exhibit regions where there has been insufficient observations (see the activity maps) to resolve the depth structure. In particular, the depth of furthest regions is determined by the deepest observations. In addition, the clock signal appears as a nearby scene element.

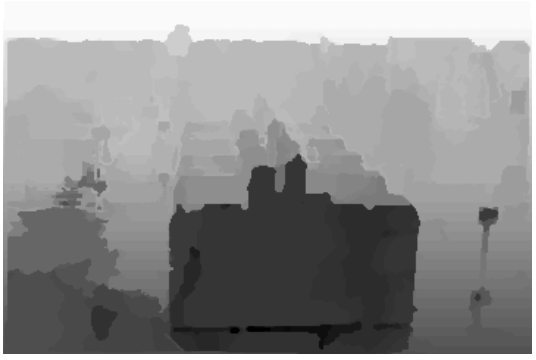

Fig. 3. Regularised Depth Map

3. EVALUATION

The utility of the depth structure in improving tracking accuracy is best demonstrated by a specific example of a figure traversing the scene and being occluded over 80 frames (or three seconds) by the turnstiles - see Figure 4(a).

Fig. 4. (a) Detected occluded object (b) Template

New locations of the object are recovered using the template correlation *i.e.* the sum of squared greylevel differences between template and image pixels[6]. (The template of the tracked object of Figure 4(a) is shown in Figure 4(b)). In this approach, correlation is initially applied to all pixels which belong to the template irrespective of whether these pixels have been occluded or not. However, the depth map can be used to identify those template pixels which are likely to be occluded by comparing the predicted depth of the object (from the template's predicted image position) with the depth value of the scene.

Figure 5 plots match probability results for correlation[1] both with and without this occlusion analysis. The first plot shows the correlation result without any identification of the occluding pixels while in the second plot, the depth map has been used to identify occluding pixels. Without any occlusion analysis, the match probability drops dramatically as the occluding surface significantly distorts the appearance of the object. However, when using occlusion analysis, the match probability remains high throughout the occlusion.

4. CONCLUSION

Static occlusions are a significant source of failure in object tracking as they significantly alter the expected appearance of the object. Depth masks can play an important role in predicting how static occlusions affect the expected appearance of an object, and hence, aid the data association. A method of generating the probability density functions (PDFs) of the likely depth of the scene at each pixel is presented. This learning approach uses a training set of observations of detected moving people, each of which constrains part of the occluded scene to lie at some distance beyond the observation. Since the results tend to be noisy, a regularisation process is required. Occlusion boundaries generate discontinuities in the activity map which can be used to prevent the smoothing of depth over possible depth boundaries. Having extracted the depth scene, we have illustrated how

[1]The calculation of the probability assumes that correlation values belong to a Chi-squared distribution whose number of degrees of freedom is given by the number of pixel comparisons used in the correlation.

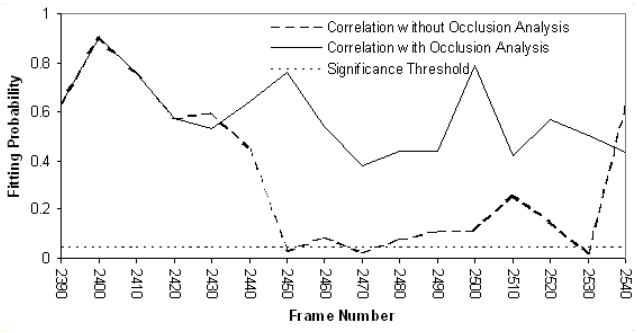

Fig. 5. Correlation probabilities across occlusion

the depth map can aid the inter-frame correspondence problem which is so highly sensitive to occlusion.

Acknowledgment

This work was sponsored by the Engineering and Physical Sciences Research Council on grant GR/S98443/01 *REVEAL: Recovering Evidence from Video by fusing Video Evidence Thesaurus and Video Meta-Data*.

5. REFERENCES

[1] T. Aach and A. Kaup. "Bayesian algorithms for adaptive change detection in image sequences using Markov random fields". *Signal Processing: Image Communication*, 7(2):147–160, August 1995.

[2] T. Ellis and M. Xu. "Object Detection and Tracking in an Open and Dynamic World". In *Second IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, Hawaii, December 2001.

[3] N. Paragios and G. Tziritas. "Adaptive Detection and Localization of Moving Objects in Image Sequences". *Signal Processing: Image Communication*, 14:277–296, September 1999.

[4] J. Renno, J. Orwell, and G.A. Jones. "Learning Surveillance Tracking Models for the Self-Calibrated Ground Plane". In *British Machine Vision Conference*, pages 607–616, Cardiff, UK, September 2002.

[5] A. Schodl and I. Essa. "Depth Layers from Occlusions". In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 639–644, Kawai, Hawaii, December 2001.

[6] Andrew Senior. "Tracking People with Probabilistic Appearance Models". In *Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 48–55, Copenhagen, June 1 2002.