Probabilistic Parameter Selection for Learning Scene Structure from Video

M. D. Breitenstein[1] E. Sommerlade[2] B. Leibe[1] L. Van Gool[1] I. Reid[2]
[1]ETH Zurich, Switzerland [2]University of Oxford, UK

Abstract

We present an online learning approach for robustly combining unreliable observations from a pedestrian detector to estimate the rough 3D scene geometry from video sequences of a static camera. Our approach is based on an entropy modelling framework, which allows to simultaneously adapt the detector parameters, such that the expected information gain about the scene structure is maximised. As a result, our approach automatically restricts the detector scale range for each image region as the estimation results become more confident, thus improving detector run-time and limiting false positives.

1 Introduction

Video surveillance is becoming a major application area of computer vision. New cameras are installed daily all around the world, adding huge quantities of data to the streams of information that need to be processed. As this happens, it becomes increasingly important to develop methods for reducing the manual effort that is still required for camera setup, replacing it by automatic processing.

In this paper, we focus on one aspect of this problem, namely to learn as much as possible about the depicted scene in an entirely automatic fashion. Our goal is to learn the rough scene geometry and semantics (e.g. where can people walk? Where do they typically appear at which sizes?) from videos of a static camera by observing objects that move within it. Taking advantage of the fact that observed object size is related to distance, we accumulate responses from a pedestrian detector to infer 3D structure. As the output of a pedestrian detector however still contains considerable noise and many false detections, especially for the difficult video footage available from many surveillance settings (see Fig. 1), we continuously integrate those measurements in a graphical model for robust estimation. In return, we use the estimated scene structure to constrain the object detector only to those image regions and local scales at which objects are likely to occur (see Fig. 2). Such a procedure can result in considerable speedups, which go hand-in-hand with an increase in accuracy, since false positives at improbable scales are already filtered out (see e.g. [10, 14]).

Several other approaches have been proposed recently which target a similar goal by trying to estimate scene structure in a batch learning stage (e.g. [13, 16]). For an automatic system, such a fixed learning stage may be problematic, since it is not guaranteed that the limited number of observations during that stage provides sufficient information for defining the entire scene. Consequently, those approaches employ additional simplifying assumptions about camera viewpoints [9, 17], specific properties of indoor scenes [6] or urban scenes [3, 18], or the existence of one single ground plane that is valid for the entire scene [10, 13, 16]. In contrast, our approach applies the learning process in an *online*

BMVC 2008 doi:10.5244/C.22.32

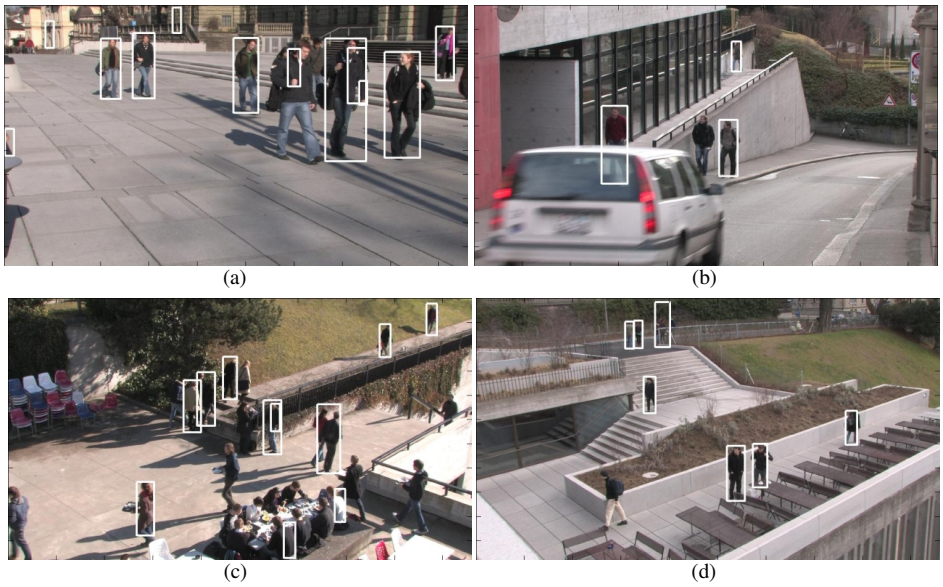
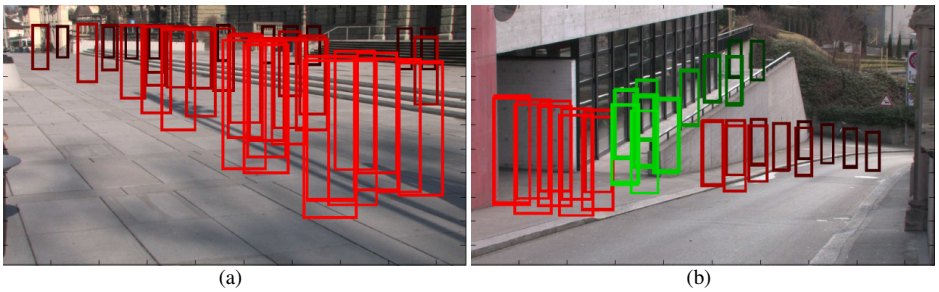


Figure 1: The typical output of a pedestrian detector (here, HoG [5] is used) on complex scenes (containing shadows, occlusions, and multiple ground surfaces), with many false positives, missing detections, and inexact bounding boxes. Our target is to robustly estimate multiple "walkable" ground surfaces from such noisy observations.

fashion such that the estimation can be refined after each observed frame. Furthermore, it is able to deal with scenes that consist of arbitrary (possibly several) ground surfaces.

Since the online estimation is based on incomplete information, this raises an interesting problem: For computational reasons and to filter out false positive detections, we want to restrict the search space for location and size of pedestrians in the scene as far as possible. However, although we may not have observed a person in a certain image region so far, this does not mean that one cannot appear there, requiring us to continuously re-explore the search space. To resolve this tradeoff, we propose a method for informed parameter selection which minimises the expected uncertainty of the scene structure estimate based on an entropy framework. Our approach allows to gradually restrict the scale range parameter of a pedestrian detector, as the scene estimation becomes more reliable. If no information about the scene is available, all possible observations are considered (exploration phase), and the more reliable the estimation becomes, the more specific the chosen search scale ranges get (exploitation phase). Fig. 2 shows samples for a location-dependent size prior obtained after observing about 30s of a video sequence.

In detail, this paper makes the following contributions. 1) We propose a method to learn the local geometry of a scene from the output of a pedestrian detector by estimating multiple "walkable" ground surfaces. 2) This method is based on an entropy framework to simultaneously and gradually adapt detector parameters (in our case the scale range parameter) such that the expected information gain about the scene structure is maximised. 3) Both of those steps are combined in an online learning approach to benefit from continuous data sources, while maintaining quality-of-service for the resulting detections.



(a) (b)

Figure 2: Our method automatically estimates "walkable" ground surfaces from which location-dependent size priors for pedestrian detection can be obtained. Here, randomly sampled bounding boxes are shown for two scenes. The colors depend on the ground surfaces (two are found in 1(b)), and the intensity is proportional to the estimated depth.

2 Related Work

3D Reconstruction. An explicit 3D reconstruction of the scene requires two or more images taken from different viewing locations with overlapping field-of-view [8]. For many surveillance settings, this is not practical, as most data sources are static and monocular cameras. Single-view scene geometry estimation has therefore attracted increasing interest recently. However, existing approaches are usually based on strong assumptions about the scene or work only in specific settings: Shape-from-shading [20] is only applicable to scenes of uniformly coloured and textured surfaces, and methods based on 3D metrology [4] require manual interaction. Other attempts are based on the assumption of orthogonal shapes in the scene, which are predominant in urban environments [3, 18, 11] or indoor scenes [6]. Usually, these methods fail for general outdoor settings. One of the most popular methods [9] assume that the world consists of vertical structures and one flat ground surface, and make strong assumptions about the viewpoint (i.e. camera position and axis). Then, a classifier is learnt to model the relation between local material properties (colour and texture), 3D orientation, and image location. In [17], this approach is improved such that it works for non-vertical structures. In contrast to these methods, our algorithm is based on integrating pedestrian detections over time instead of directly using geometric features. Therefore, it does not dependent on any assumptions about scene geometry or viewpoint. Furthermore, our algorithm can learn multiple ground surfaces. However, it is restricted to reconstruct ground surfaces where people walk. Our algorithm is based on an online learning process to benefit from virtually unlimited data sources like web-cams or surveillance cameras.

Camera Autocalibration. For surveillance applications, camera autocalibration has become a popular tool, closely related to 3D scene reconstruction. A possible approach is to estimate vanishing directions through edge detection and grouping (e.g. [11]). Although the 3D orientation of a plane can be determined by its vanishing line (relative to the camera) [8], such information cannot easily be extracted from unstructured outdoor images and is notoriously sensitive to noise. Another approach is to exploit homologies obtained from moving pedestrians in the scene. However, in previous work [13, 12, 16] head and foot positions of single pedestrians needed to be detected and tracked very accurately

and robustly. Although also based on moving pedestrians, our approach is not dependent on reliable tracking. Instead, it pursues a conservative online learning strategy by only picking out confident detections (e.g. from image regions where pedestrians are walking individually) and integrating those over a longer time window. Therefore, our method is more robust and can be used for complex and crowded scenes.

Improving Object Detection. Recent work has shown how scene knowledge can be used to improve object detection by providing contextual priors for object location and scale [10, 19]. For example, the viewpoint can be modelled explicitely in a graphical model to combine object detection and geometric context [10]. In contrast to previous methods, our approach is to gradually adapt detector parameters online by maximising the expected information gain for the scene estimation. Therefore, we optimally reduce the uncertainty in the state estimate by choosing the best observation parameters. Such an approach has previously been used for control parameter optimisation for active sensors (such as pan, tilt and zoom cameras) [7]. However, while previous work is concerned with optimising the actual parameters of a physical device, we apply this idea to parameter selection for the image-based detection process. To our knowledge, no prior work exists in this area.

3 Approach

Our method estimates the scene structure based on noisy observations of pedestrians, and simultaneously and gradually adapts the detector parameters such that the uncertainty about the scene estimation is minimised. We consider observations about objects in a scene as sequential time-series data and model the dependencies between noisy and unreliable observations and the scene structure at a certain position in the image in a graphical model. The observations stem from a sliding-window based HoG pedestrian detector [5]. However, our approach is independent of the specific detector.

We divide the image into a regular grid (chosen independently from the specific scene), and estimate the depth and spatial structure for each cell independently in a Dynamical Bayesian Network. The depth estimation is a probability distribution over relative depth classes, and the spatial structure is the probability that pedestrians appear ("walkability"), which provides an indication for the reliability of the estimated depth.

Based on these online estimations, the control parameters for the detection algorithm are adapted continuously for each image region, such that the expected information gain is maximised in each time step. Equivalently, this corresponds to a minimisation of the expected uncertainty, which we model by the conditional entropy of the scene estimation. By optimising the detections and thus the scene structure estimation, we get more reliable measurements in the next time step. An essential parameter of an object detector is the search scale range, which defines the size of pedestrians to look for. We restrict the optimisation to find such *location-dependent scale range parameters*. Essentially, one parameter is a single pointer to a table of scale ranges, defining the minimum and maximum scale of objects the detector is looking for (see Fig. 5(c) for a legend of the scales covered by the pointers in our experiments). However, our approach could also be applied to adapt other parameters.

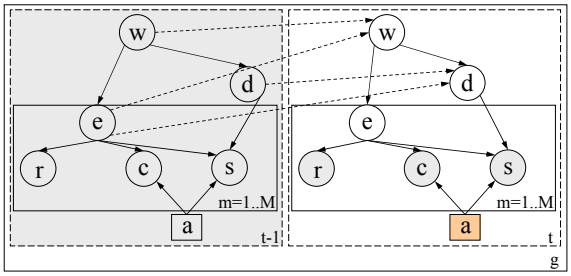


Figure 3: The graphical model for one image cell g and one time step t. The observed nodes s, c are the measurements from the pedestrian detector (*scale*, *confidence value*) and r stems from data association (*reliability*). The latent nodes e, d, w correspond to the estimated *existence of an object*, and the *depth* and *walkability* estimations. The control (or *action*) parameter a is the selected scale range parameter of the detector.

3.1 Graphical Model

At the core of our approach is a Dynamic Bayesian Network (DBN) for inferring scene structure from object detections. Each DBN models one cell of the image grid (see Fig. 3). Its purpose is to integrate information from object detector responses in order to simultaneously estimate local scene structure and to provide the entropy framework with a means of predicting what effects a change in the detector parameters will have on the expected information gain. The object detector yields observations with a confidence and scale measurement, which are added as evidence into our graphical model.

Observed Variables. For each image cell g, we add m measurements from the current time step t. The multinomial variable $s \in \{s_1, \ldots, s_N\}$ corresponds to the *scale* of a detection. Furthermore, the binary variable $c \in \{0, 1\}$ is the *confidence value* of a detection after thresholding. Thus, if the confidence value obtained from the detector is low, the measurement will have less impact than if the confidence value is high. In addition, we check whether the observed detection has consistently moved compared to the preceding time steps. This is necessary in order to filter out certain background image structures, such as doors or windows, which give rise to very persistent false positives in some image locations. The result of this data association check is added through an observed binary variable $r \in \{0, 1\}$ (*reliability*), which ensures that a pedestrian is moving, thus eliminating potentially wrong static detections. As a result, our procedure also ignores possibly correct detections of standing pedestrians when estimating scene geometry, which may lead to slightly longer time until convergence for the benefit of added robustness (against false, early conclusions due to possibly wrong detections).

Latent Variables. Our target is to reliably estimate the depth values for "walkable" regions which best explains the measurements. For each time step and each measurement m, the evidence (r, c, s) is added to the model, and the probability of the *existence of an object* e for the current time step in this cell is inferred. Hence, the latent binary variable $e \in \{0, 1\}$ combines all the indications from the observed nodes (r, c, s) to estimate whether a reliable observation has been made at the current time step. The latent variable $w \in \{0, 1\}$ corresponds to the estimation, how likely an object will appear at this image location, and serves as a prior for e. Because we use measurements from a pedestrian de-

tector, and pedestrians only appear in scene regions which are "walkable", w corresponds to the *"walkability"* of this cell. Finally, the relative *depth* of the scene is modelled by the latent multinomial variable $d \in \{1,\ldots,D\}$. The estimated depth d depends on the walkability w, because we only can reliably estimate the depth for image regions where it is likely that pedestrians can appear. The control (or *action*) parameter a models the influence of the detector parameters (in our case the search scale range) on the output of the detector. a is selected by our algorithm (see Section 3.2) based on the last time step and fixed for the current time step t, thus not taken into account for inference. Ultimately, we want to optimise the parameters a so as to minimise the expected uncertainty of the depth estimate for walkable regions $p(d, w = 1)$.

Structure. The dependencies of the variables are encoded in the DBN, see Fig. 3. The conditional probability distributions of the variables within one time slice are manually chosen based on experiments (independently from the test videos), and the behaviour of the DBN is verified for synthetic test sequences.

The current depth and walkability estimations depend on the last time step. The number of observations (based on which the estimations are obtained) influences the reliability of the estimations. Therefore, if only few observations have been made so far, a new and reliable observation should influence the estimation stronger than if many reliable observations already have been made before. We model this in a two-slice Temporal Bayesian Network (2TBN) (a special case of a DBN, [15]) by the transition probabilities $p(w_t|w_{t-1}, e_{t-1})$ and $p(d_t|w_t, d_{t-1}, e_{t-1})$ (corresponding to the dashed lines in Fig. 3). They are modelled by unnormalised histograms and updated if a successful observation was made (i.e. if $p(e_{t-1} = 1)$). Hence, the a priori probabilities for walkability and depth depend on the old estimation, and are updated if a reliable observation has been made. With these modifications, we add a "memory" to the graphical model.

The updated conditional probability distribution in the 2TBN can be computed from the inferred conditional probability distribution in the current time slice and the transition probabilities introduced before:

$$p_a(d_t, w_t, e_t|d_{t-1}, w_{t-1}, e_{t-1}, c_t, s_t, r_t) =$$
$$p_a(d_t, w_t, e_t|c_t, s_t, r_t)p(d_t, w_t, e_t|d_{t-1}, w_{t-1}, e_{t-1}) = \quad (1)$$
$$p_a(d_t, w_t, e_t|c_t, s_t, r_t)p(w_t|w_{t-1}, e_{t-1})p(d_t|w_t, d_{t-1}, e_{t-1})p(e_t|w_t) \quad (2)$$

For inference in the 2TBN, we use an approximation of the junction tree algorithm implemented using forward/backward operators (see [1]).

3.2 Entropy-based Optimisation of Detector Parameters

We aim at adapting the detector parameters in each time step depending on the estimated scene structure, i.e. the depth estimate for walkable regions $p(d, w = 1)$. Therefore, we optimally reduce the expected uncertainty in the state estimate by choosing the best observation parameters. This influences the observations and thus the inferred scene estimation. The control parameter a_t summarises the different parameter settings for the observation process, in our case the possible scale range parameters of the detection algorithm.

At time step t, our algorithm chooses the best possible parameter a_t to make an observation $\mathbf{o}_t$ based on the last scene estimation $p_{t-1}(w, d|a)$. Among all possible choices, the selected parameter will maximally reduce the expected uncertainty in a given probability

distribution of the true state $\mathbf{x} = (w,d)$ of the DBN, which is the scene estimation we want to optimise. Applying the chosen parameter a_t yields an observation $\mathbf{o}_t = (c,s,r)$ which is finally used to update the scene estimation $p_t(w,d)$ by Bayesian inference in the DBN.

A natural measure for the uncertainty is the expected conditional entropy,[1]

$$\hat{H}_{a_t}(\underbrace{w_t,d_t}_{\mathbf{x}_t} \,|\, \underbrace{e_t,c_t,s_t,r_t}_{\mathbf{o}_t}) = -\iint p_{a_t}(\mathbf{x}_t,\mathbf{o}_t)\log(p_{a_t}(\mathbf{x}_t|\mathbf{o}_t))\,d\mathbf{x}_t\,d\mathbf{o}_t \tag{3}$$

which takes into account the distribution of the observations, and hence is independent of the observation to make in the next time step. The parameter $\mathbf{a}_t^*$ is then found by minimising the entropy:

$$\mathbf{a}_t^* = \arg\min_{\mathbf{a}_t} \hat{H}_{\mathbf{a}_t}(\mathbf{x}_t|\mathbf{o}_t) \tag{4}$$

The probabilities in Eq. 3 are obtained by inference in the DBN (see Section 3.1), that encodes the conditional independence relationships, and by applying the chain rule of probability. Thus, we can write the probability of all nodes at time instant t:[2]

$$p_a(w,d,e,c,s,r) = p(w)p(d|w)p(e|w)p(r|e)p_a(c|e)p_a(s|e,d) \tag{5}$$

The factors influenced by the detector are $p_a(c|e)$ and $p_a(s|e,d)$, which need to be estimated for every possible parameter setting a (see next paragraph).

Our approach results in a restriction in the data selection process: We reduce the scale range of possible targets in the observed scene and thus implicitly apply a location-dependent size prior. As a result, our method decreases the number of false positives in areas where it can acquire sufficient detections to reliably estimate scene geometry.

Training: Estimation of Detector Statistics. The influence of the detector parameters on the measurement process can be summarised by the probabilities $p_a(c|e)$ and $p_a(s|e,d)$. They describe the likelihood of making an observation of a certain confidence c given the evidence of an object class e, and of size s given e and the depth d. Usually, neither true depth d nor true height of a person are found in publicly available image databases. Instead, only the pixel size s^* of persons in the image is annotated. Therefore, we learn the output of the detector for the annotated pixel size s^* from training data, yielding $p(c|e,s^*)$ and $p(s|e,s^*)$, and approximate $p(c|e)$ and $p(s|e,d)$ by marginalisation:

$$p(c|e) = \int p(c|e,s^*)p(s^*|e)\,ds^* \tag{6}$$

$$p(s|e,d) = \int p(s|e,s^*)p(s^*|e,d)\,ds^* \tag{7}$$

Furthermore, we model $p(s^*|e,d)$ with a projective mapping for an assumed focal length f and object height h:

$$p(s^*|e=1,d) = k\exp(-(fh/d-s^*)^2/\sigma^2) \tag{8}$$

and estimate the false detection distribution $p(s^*|e=0,d)$ by running the detector on images without pedestrians. $p(s^*|e=1)$ is the distribution of annotated object sizes in

[1] $p_a(\cdot) = p(\cdot|a)$ is a probability that additionally depends on the control parameter a.

[2] To simplify the notation, we omit the subscript t and marginalise out the variables at $t-1$.

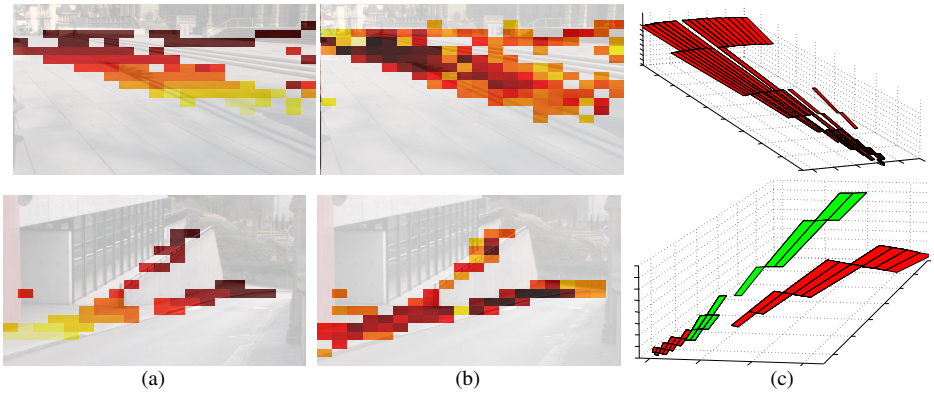


Figure 4: Results for the scenes in Figs. 1(a) and 1(b). a) most probable depth classes for walkable regions (depth encoded in colour values). b) probability of walkability (i.e. probability of reliably detected pedestrians; dark regions are more probable). c) ground surfaces fitted to the depth estimates. (best viewed in colour)

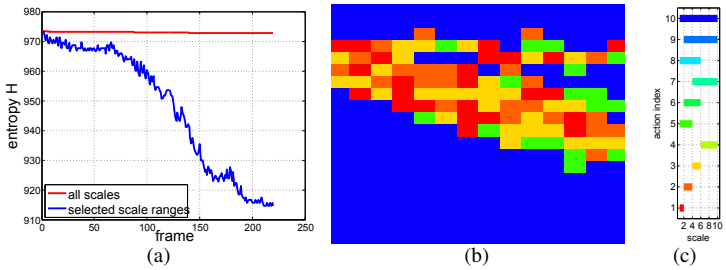
the training set. For training the expected detector performance, we used the annotated GRAZ/INRIA image database, and the detection tolerances from [2]. Although Eq. 8 neglects the actual focal length of the original images, manual inspection of the data set used for the evaluation reveals a typical short focal length for street scenes, which coincides with our application domain.

4 Experimental Results

We ran several experiments to verify our method. For all tests, the frames from high resolution (HD) videos are discretised into a regular 20×20 grid, each cell containing a DBN as described before. The number of depth classes d and the number of different scale range parameters a are set to 10. We divide the maximum search scale range for the detector in 4 parts, and map the possible minimum/maximum scale combinations to 10 different actions (see the legend in Fig. 5(c)).

In Fig. 4, we show results of the scene structure estimation after about 30 seconds for videos from the screenshots in Figs. 1(a) and 1(b). Column a) shows the most probable depth classes $\max_d p(d, w=1)$ where pedestrians are reliably detected (i.e. "walkable" regions), and column b) shows the probability that a region is walkable $p(w=1)$. Although the detection algorithm returns many false positives for such cluttered and crowded high-resolution images with many partially occluded pedestrians, the DBN robustly estimates the regions of reliable detections and the relative depth classes. If running for a longer time, more measurements will be combined subsequently by the online algorithm. Therefore, the depth and walkability estimates will get more reliable as the estimated distributions get narrower. Because more measurements fall into one cell for regions which are far away, it usually takes less time until the estimation converges (see e.g. Fig. 4(b), top). Furthermore, pedestrians probably will appear at locations which were not observed before, resulting in a more complete scene estimation.

To estimate the ground surfaces which best fit the estimated depth values, we project



(a) (b) (c)

Figure 5: a) The entropy of the scene estimation with and without scale range parameter selection. b) Location-dependent search scale ranges for each cell after 30s. c) Colour legend for the action parameter a, which points to a range of detector search scales.

the depth values to 3d points by roughly assuming a focal length of the camera (which is either known or a rough guess is estimated easily, c.f. Section 3.2). Planes are then automatically fitted to the points using RANSAC and segmented into walkable regions where $p(w = 1)$ (results are shown in Fig. 4(c)). For the sequence in the 2nd row, two ground surfaces are correctly found. Such a result was not possible with previous methods.

Secondly, we demonstrate the effect of the entropy-based search scale range selection in an online experiment. In each frame, measurements from the detector are combined, and the best location-dependent parameters for the next time step are selected. In the beginning, the depth distribution for one cell is rather uniform. After each time step, our method progressively refines it, and its entropy decreases. In Fig. 5(a), the entropy of the scene estimation is plotted for each frame with and without our algorithm for scale range parameter selection. In Fig. 5(b), the chosen parameters for each cell after about 30s are shown for the scene in Fig. 1(a). Fig. 5(c) illustrates which detector search scale range is covered by which action parameter a, encoded in colours. Fig. 5(b) demonstrates that the search scale range is strongly restricted in cells with reliable scene estimates (c.f. Fig. 4, top). For image regions where the scene estimation is not reliable (because no or noisy observations have been made so far), the search scale range is less limited.

For a typical frame of our HD video sequences, the original pedestrian detector [5] evaluates more than 300′000 detection windows on the whole image. In comparison, our method effectively limits the search scale range such that typically only about 40′000 windows are evaluated for reliably estimated image regions (a 10x speedup). For the remaining region, one could employ a random sampling scheme with a small budget of detection windows, since the region is less likely to contain pedestrians. Such methods to maintain quality-of-service for detections are especially important for online applications like surveillance.

5 Conclusion

We presented an online learning approach to estimate the local geometry of the scene based on unreliable observations from a pedestrian detector. Our method is not based on strong assumptions (e.g. about the viewpoint), can learn multiple ground surfaces, and benefits from virtually unlimited data sources like web-cams or surveillance cameras. Simultaneously, the algorithm adapts the scale range parameter of the detector such that the

expected information gain about the scene structure is maximised, allowing to improve the detector run-time and limiting false positives. Hence, our approach resolves the trade-off of exploring the image for all possible scales vs. relying on the current estimation. To our knowledge, no previous work exists on online learning of detection parameters based on entropy. Our algorithm is currently restricted to reconstruct ground surfaces where people walk. In the future, we plan to extend the graphical model to include conditions between cells, and to additionally use colour and texture cues.

Acknowledgments: We thank A. Ess and V. Ferrari for help and valuable comments. The authors gratefully acknowledge support by the EU project HERMES (IST-027110).

References

[1] http://www.cs.ubc.ca/ murphyk/Software/BNT/usage-dbn.html.

[2] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI*, 26(11), 2004.

[3] J. M. Coughlan and A. L. Yuille. Manhattan world: Orientation and outlier detection by bayesian inference. *Neural Comput.*, 15(5), 2003.

[4] A. Criminisi, I.D. Reid, and A. Zisserman. Single view metrology. *IJCV*, 40(2), 2000.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR'05*.

[6] E. Delage, H. Lee, and A. Ng. A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In *CVPR'06*.

[7] J. Denzler and C. M. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *PAMI*, 24(2), 2002.

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. 2004.

[9] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV'05*.

[10] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR'06*.

[11] J. Kosecka and W. Zhang. Video compass. In *ECCV'02*.

[12] N. Krahnstoever and P.R.S. Mendonca. Autocalibration from tracks of walking people. In *BMVC'06*.

[13] N. Krahnstoever and P.R.S. Mendonca. Bayesian autocalibration for surveillance. *ICCV'05*.

[14] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool. Dynamic 3d scene analysis from a moving vehicle. In *CVPR'07*, 2007.

[15] K. Murphy. Dynamic bayesian networks. In *Probabilistic Graphical Models*. 2002.

[16] D. Rother, K. A. Patwardhan, and G. Sapiro. What can casual walkers tell us about the 3d scene? In *ICCV'07*.

[17] A. Saxena, M. Sun, and A. Ng. Learning 3-d scene structure from a single still image. In *ICCV Workshop on 3d Representation for Recognition (3dRR'07)*.

[18] G. Schindler and F. Dellaert. Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. *CVPR'04*.

[19] A. Torralba and P. Sinha. Statistical context priming for object detection. In *ICCV'01*.

[20] R. Zhang, P. Tsai, J. Cryer, and M. Shah. Shape from shading: A survey. *PAMI*, 21(8), 1999.