# DEPTH ESTIMATION VIA STAGE CLASSIFICATION

*Vladimir Nedović* [†], *Arnold W.M. Smeulders* [†], *André Redert* [‡] *and Jan-Mark Geusebroek* [†]

[†] Intelligent Systems Lab Amsterdam (ISLA), University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

[‡] Philips Research Laboratories Eindhoven
High Tech Campus 36, 5656 AE Eindhoven, The Netherlands

## ABSTRACT

We identify scene categorization as the first step towards efficient and robust depth estimation from single images. Categorizing the scene into one of the geometric classes greatly reduces the possibilities in subsequent phases. To that end, we introduce 15 typical 3D scene geometries, called *stages*, each having a unique depth profile and roughly corresponding to a large majority of all images. In this work, we do not attempt to derive a precise depth map, but only to decide on the appropriate stage. The subsequent phase of parameter estimation would result in a more detailed background depth profile.

***Index Terms***— Depth estimation, 2D to 3D conversion, scene geometry, scene classification, surface layout, stages

## 1. INTRODUCTION

Extracting depth from single images is important for many applications, including three-dimensional television (3DTV). The images themselves can be separated into two constituent parts that require different treatment in this process: the physical scene and the objects acting in it. The objects of the world come with almost infinite variation in appearance as well as in their geometry. Scenes, on the other hand, show a much more regular pattern. In fact, the vast majority of images depicts a scene by only a limited number of different geometry types. There are rough classes of such geometries, or *stages* as we prefer to call them, which consist of a straight background (like a curtain, a wall, the façade of a building, a remote mountain range), or others showing walls at all three sides of the picture (a corridor, a tunnel, a narrow street). When television broadcasts are considered, there is also a specific stage for anchor-type images, corresponding to news-reader sequences, interviews, talk-shows, etc. Figure 1 shows a few prototypes together with their stage models. The important observation is that whereas a precise geometry is requested for objects, it suffices to build a rough model for the geometry of the scene. The structure of scenes poses the question of this paper. More specifically, we aim to discover whether stages as models of scene geometry can be derived from a single arbitrary image.

There is a good reason why the geometry of the scene can be represented by one of very few classes. Human observers almost always stand with their feet on the ground, walls are almost always perpendicular to the ground, they are to the side of the object or behind it, and so on. Moreover, there is an advantage of knowing just the stage type. The stage may reveal to the observer the type of the scene, the information about relative distances to scene elements, the locations in the field of view where objects may appear, etc.

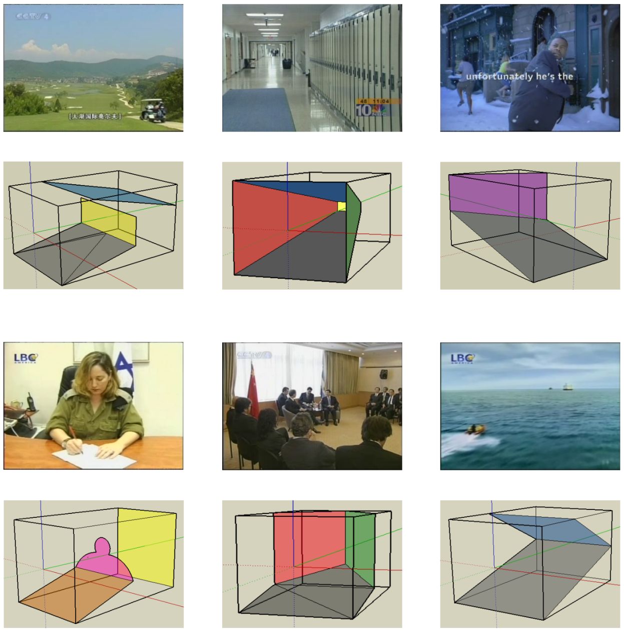

**Fig. 1**. Example frames and their stage categories; top two rows, from left to right: *sky+ground, table+person+background, diagonal background*; bottom two rows: *box, corner, sky+background+ground.*

We are targeting an application in 3D television, and thus we aim for a pleasant 3D viewing impression rather than a precise reconstruction of scene geometry. Accurate techniques have been designed for object geometry via shape from shading, shape from motion or shape from stereo, when these options are available. The scene geometry, in contrast, is the stage on which the objects of the picture act, hence limited accuracy frequently suffices. In this paper, we consider stages as very rough models of the scene, with the objects ignored.

In the presented work, we limit ourselves to the determination of the stage type, following closely the work from [1], but with an emphasis on 3DTV applications. In the next phase, further depth estimation can be performed, by estimating the stage parameters from the data. Here we present stage classification results for the domain

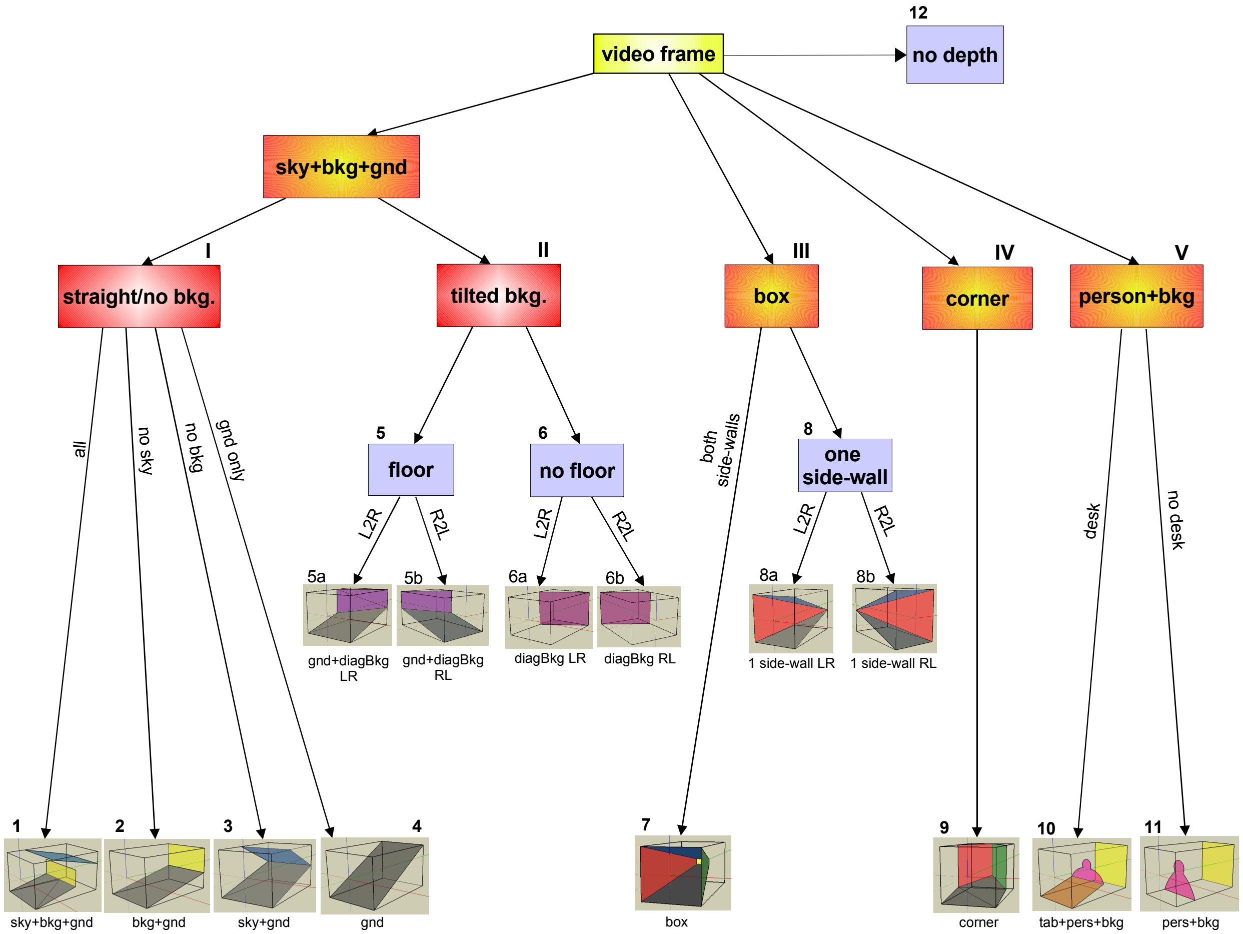

**Fig. 2**. Stage hierarchy: the classes at the intermediate level are represented by Roman numerals I-V, whereas those at the lowest level are represented by Arabic numerals 1 through 12. Note that the symmetry of certain stages is represented by an additional division of the stage type.

of TRECVID news videos [2].

## 2. RELATED WORK

Many recent attempts to estimate absolute scene depth from single images use machine learning methods to directly infer image depth from simple features. Torralba and Oliva [3] use global Fourier transform and local wavelet transforms to capture scene structure; Saxena et al. [4] use features of multiple depth cues; and Delage et al. [5] attempt to learn the wall-ground boundaries.

For convincing visual 3D quality, derivation of exact distances to elements in the scene may not be necessary as long as relative order of those elements is established. However, classical methods for relative depth estimation provide only local estimates and require high-quality images, as is the case for texture gradients [6], shape from shading (e.g. [7]), from edges and junctions [8], etc.

Scene classification approaches [9, 10, 11, 12, 13] attempt to capture the complex statistics of natural images. However, they suffer from two drawbacks that render them unsuitable for depth estimation. The first is that they model semantic scene categories (such as kitchen, forest, street, etc.), whose potential number is very large. The second drawback is that all the approaches consider only the 2D image, without attempting to recover the 3D scene. To that regard, *geometric* image context has recently been used instead by Hoiem et al. [14, 15]. They learn classes of image surfaces and derive the orientation of each such class; the subsequent combination of surface orientations leads to the reconstruction of the 3D scene model.

### 2.1. Contribution of the paper

We draw inspiration from the work of Hoiem et al. [14] and attempt to derive the 3D geometry of the scene, and recover depth information. However, instead of individual surfaces, we model geometric scene classes. We believe that the recognition of the scene as a whole into a limited number of typical stages is a simpler problem than image segmentation and subsequent reconstruction.

Our work on depth estimation is also similar to that of Torralba and Oliva [3], who showed that depth can also be derived from models of natural image statistics. But where they propose to utilize average absolute depth in order to facilitate scene categorization, we attempt to do the opposite, and propose to derive a global depth profile based on stage types. In addition to relying on natural image statistics, we take into account the viewpoint of the observer. This greatly reduces the number of categories that we need to model.

## 3. STAGE TYPE CLASSIFICATION

We rely on the structure of the visual data to arrive at a limited number of stage types. This structure is a consequence of three phenomena: natural images exhibit statistical regularities; viewpoint characteristics (including the camera height typically at $1.5 - 2m$) govern the perspective; and film rules ensure the proper order of elements (e.g. that ground is at the bottom) and the "uprightness" of image structures.

We have looked at thousands of TRECVID keyframes [2] and noted the frequency with which various scene categories appear. The structure that we observed limited all possible surface combinations to 15 categories only. Retaining only the stage types corresponding to more than 5% of the observed video frames, they were sufficient for a large majority of all the frames.

When one represents semantic scene classes, the natural top-level categorization is into indoor and outdoor images. However, when geometrical classes are considered, the 3D constraints (such as e.g. perspective) cause certain stage types to be represented by the same 3D model, *regardless* of whether they belong to indoor or outdoor scenes. Thus we have concluded that stage hierarchy should be based on geometry and arrived at the representation shown in Figure 2.

### 3.1. Depth from Weibull texture gradients

There exists a direct relation between image statistics, scene structure and depth pattern. When scene depth is small, larger surfaces merge into coarser structures, showing finer details. In that case, gradient histogram typically follows a decaying power-law distribution. When scene depth increases, the texture of the image will be fragmented into various patches, each associated with a different power-law. The integration over various power-laws results in a Weibull distribution [16], whose parameters are indicative of depth direction. This is shown in Figure 3 for two example surfaces.

We follow [16] and model histograms of gradient magnitude by an integrated Weibull distribution, also known as Generalized Laplacian,

$$f(x) = \frac{\gamma}{2\gamma^{\frac{1}{\gamma}}\beta\Gamma\left(\frac{1}{\gamma}\right)} e^{-\frac{1}{\gamma}\left|\frac{(x-\mu)}{\beta}\right|^{\gamma}} \tag{1}$$

The parameters $\mu, \beta$ and $\gamma$ represent the center, width and "peakness" of the distribution, respectively, and $x$ is an edge response of a derivative filter. Furthermore, $\Gamma$ denotes the complete Gamma function.

Using Gaussian derivative filters, we extract texture information that is subsequently summarized in histograms. We use a maximum likelihood estimator (MLE) to estimate the parameters $\mu$, $\beta$ and $\gamma$ of the integral Weibull distribution. The position of the distribution mean $\mu$ is influenced by uneven illumination, and thus the values of $\mu$ are ignored in order to eliminate illumination effects.

Since our stages often contain oriented surfaces with continuously increasing depth, we perform local measurements. For each image, we extract features from consecutive image regions spanning $\frac{w}{4} \times \frac{h}{4}$, where $w$ and $h$ denote image width and height, respectively. The integral Weibull distribution is then fitted to histograms of intensity filter responses in $x$ and $y$ directions.

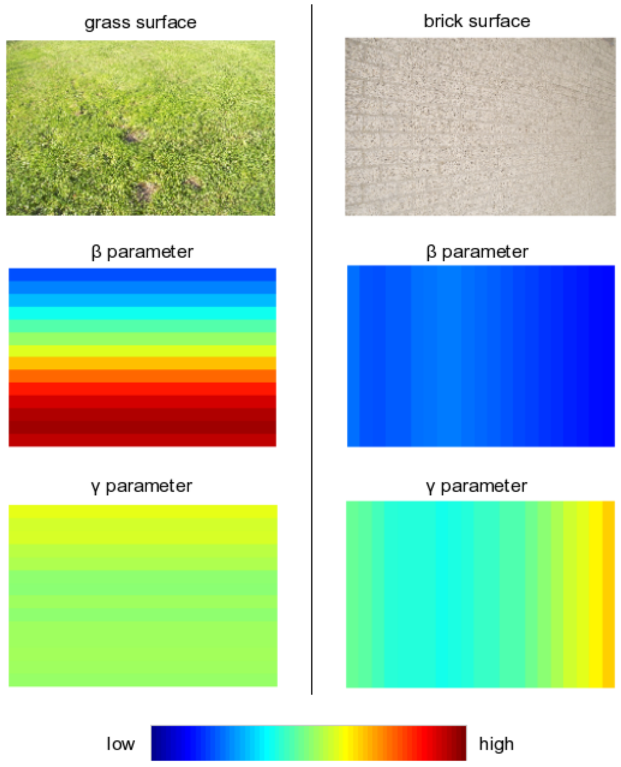

**Fig. 3**. Weibull parameter values as a function of depth for textures of grass (left) and wall bricks (right): $\beta$ decreases from the point of fixation, whereas $\gamma$ increases with depth.

## 4. EXPERIMENTS

### 4.1. Experimental setup

For the evaluation of our stage classification algorithms, we have used the key-frames of the 2006 TRECVID video benchmark dataset. This benchmark provides nearly 170 hours of news videos of various channels and languages.

In the initial result phase, we have annotated 1241 key-frames into one of the 15 stage categories. Samples of each category are split before classification into two halves, one for training and another for testing purposes. We choose Support Vector Machines (SVM) for learning, and utilize its LIBSVM implementation[1] with radial basis functions as kernels.

We design a generic, *1 vs. 1*-based classifier that uses features from all the regions and outputs a single stage label. Multi-class classifiers based on a *1 vs. 1* approach involve $K(K-1)/2$ different binary classifiers on all possible pairs of classes; test points are then classified according to which class has the highest number of 'votes'.

### 4.2. Results

Our stage classification results are shown in tables below for individual stages and stage groups, respectively. They are presented together with the relative occurrence (i.e. prior probability) of each

[1] LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/˜cjlin/libsvm

type in the ground truth. Note that, for reasons of clarity, the results for symmetrical variants of some stages have been combined. The correct classification performance is given by the total number of correctly classified images divided by the total number of images.

| *class* | *name* | *% in dataset* | *% correct* |
|---|---|---|---|
| 1 | sky+bkg+gnd | 6.3% | 16.7% |
| 2 | gnd+bkg | 7.1% | 8.2% |
| 3 | sky+gnd | 8.7% | 60.7% |
| 4 | gnd | 7.4% | 44.7% |
| 5 | gnd+diagBkg | 10.75% | 26.9% |
| 6 | diagBkg | 6.4% | 14.3% |
| 7 | box | 5.5% | 8.1% |
| 8 | 1 side-wall | 9% | 13.6% |
| 9 | corner | 10.75% | 34.3% |
| 10 | tab+pers+bkg | 7.4% | 48% |
| 11 | pers+bkg | 13.1% | 42.5% |
| 12 | no depth | 7.4% | 22.4% |
| | | | *AVG:* 28.4% |

| *group* | *name* | *% in dataset* | *% correct* |
|---|---|---|---|
| I | straight/no bkg. | 29.5% | 69.5% |
| II | tilted bkg. | 17.15% | 35.2% |
| III | box | 14.5% | 19.6% |
| IV | corner | 10.75% | 13.2% |
| V | person+bkg | 20.5% | 63.1% |
| | | | *AVG:* 40.1% |

The results indicate that some simple stages (as well as their super-stages) can be detected robustly. This is true for those classes which typically appear with small variations and without object clutter. Thus in the experiment with 12 stages, we correctly distinguish class *sky+ground* in more than 60% of the cases. On some other stages, however, our detector performance is low. This is due to the lower number of training samples, amount of variation within the class, significant occlusion and object clutter, etc. Similar observations can be made with respect to the hierarchy level with 5 stage groups. However, in all cases the performance is significantly better than chance level, indicating the usefulness of the approach.

## 5. CONCLUSIONS

In this paper, we describe how the problem of depth estimation from single images can be approached by first performing scene classification. To that end, we describe a small number of typical 3D scene geometries, or stages, each with a unique depth model and providing a geometric context for scene objects.

By relying on inherent structure of real-world images, resulting from natural image statistics, viewpoint constraints and film rules, we arrive at only 15 geometric stages (preliminary experiments on a different dataset show that these stage types are not specific to news videos). We introduce category descriptors based on natural image statistics, and show that they are indeed indicative of depth information. Quantitative classification results are presented for the data of the TRECVID 2006 benchmark, indicating that some simple stages can be detected with up to 60% success rate.

In the presented work, we do not attempt to derive a precise depth map for the input image, but only decide on the appropriate stage. For that reason, we only provide scene classification results, without comparing our models to depth map ground truth. However, the stage information helps the next phase, in which corresponding stage parameters are estimated. Once these parameters are available, a background depth map is obtained and it can be aligned with the original image in a complete depth estimation system.

## 6. REFERENCES

[1] V. Nedović, A. W. M. Smeulders, A. Redert, and J. M. Geusebroek, “Depth information by stage classification,” *ICCV*, pp. 1–8, October 2007.

[2] A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and TRECVid,” in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, 2006, pp. 321 – 330.

[3] A. Torralba and A. Oliva, “Depth estimation from image structure,” *IEEE PAMI*, vol. 24(9), pp. 1226–1238, Sep. 2002.

[4] A. Saxena, S. H. Chung, and A. Y. Ng, “Learning depth from single monocular images,” *NIPS*, 2005.

[5] E. Delage, H. Lee, and A. Y. Ng, “A dynamic bayesian network model for autonomous 3D reconstruction from a single indoor image,” *CVPR*, vol. 2, pp. 2418–2428, 2006.

[6] R. Bajcsy and L. Lieberman, “Texture gradient as a depth cue,” *Computer Graphics Image Processing*, vol. 5, pp. 52–67, 1976.

[7] T. Kanade, “Recovery of the three-dimensional shape of an object from a single view,” *Artificial Intelligence*, vol. 17, no. 1-3, pp. 409–460, 1981.

[8] H. G. Barrow and J. M. Tenenbaum, “Interpreting line drawings as three-dimensional surfaces,” *Artificial Intelligence*, vol. 17, no. 1-3, pp. 75–116, 1981.

[9] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *IJCV*, vol. 42(3), pp. 145–175, 2001.

[10] L. Fei-Fei and P. Perona, “A bayesian hierarchical model for learning natural scene categories,” *CVPR*, vol. 2, 2005.

[11] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, “Modeling scenes with local descriptors and latent aspects,” *ICCV*, vol. 1, pp. 883–890, 2005.

[12] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” *CVPR*, 2006.

[13] J. C. van Gemert, J. M. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders, “Robust scene categorization by learning image statistics in context,” in *CVPR Workshop on Semantic Learning Applications in Multimedia (SLAM ’06)*, 2006.

[14] D. Hoiem, A. A. Efros, and M. Hebert, “Geometric context from a single image,” *ICCV*, 2005.

[15] D. Hoiem, A. A. Efros, and M. Hebert, “Automatic photo pop-up,” in *ACM SIGGRAPH*, 2005.

[16] J.M. Geusebroek and A.W.M. Smeulders, “A six-stimulus theory for stochastic texture,” *IJCV*, vol. 62, pp. 7–16, 2005.