Depth Information by Stage Classification

Vladimir Nedović[1] Arnold W.M. Smeulders[1] André Redert[2]
Jan-Mark Geusebroek[1]

[1] Intelligent Systems Lab Amsterdam (ISLA), University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{vnedovic,smeulders,mark}@science.uva.nl

[2] Philips Research Laboratories Eindhoven
High Tech Campus 36, 5656 AE Eindhoven, The Netherlands
andre.redert@philips.com

Abstract

Recently, methods for estimating 3D scene geometry or absolute scene depth information from 2D image content have been proposed. However, general applicability of these methods in depth estimation may not be realizable, as inconsistencies may be introduced due to a large variety of possible pictorial content. We identify scene categorization as the first step towards efficient and robust depth estimation from single images. To that end, we describe a limited number of typical 3D scene geometries, called stages, *each having a unique depth pattern and thus providing a specific context for stage objects. This type of scene information narrows down the possibilities with respect to individual objects' locations, scales and identities. We show how these stage types can be efficiently learned and how they can lead to robust extraction of depth information. Our results indicate that stages without much variation and object clutter can be detected robustly, with up to 60% success rate.*

1. Introduction

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. The vast majority of photographs depict a scene geometry from a limited number of different types. There are rough classes of scene geometries, or *stages* as we prefer to call them, which constitute of a straight background (like a curtain, a wall, the façade of a building, a remote mountain range), or other ones which show walls at all three sides of the picture (a corridor, a tunnel, a narrow street). When video broadcasts are considered, there is also a specific stage for anchor-type images, corresponding to news-reader sequences, interviews, talk-shows, press-conferences, *etc*. Figure 1 shows a few prototypes together with their stage models. Hence, whereas in many tasks a precise geometry is requested for the object, it suffices to build a rough model for the geometry of the scene. These regularities pose the question of this paper. More specifically, we aim to discover whether stages as models of scene geometry can be derived

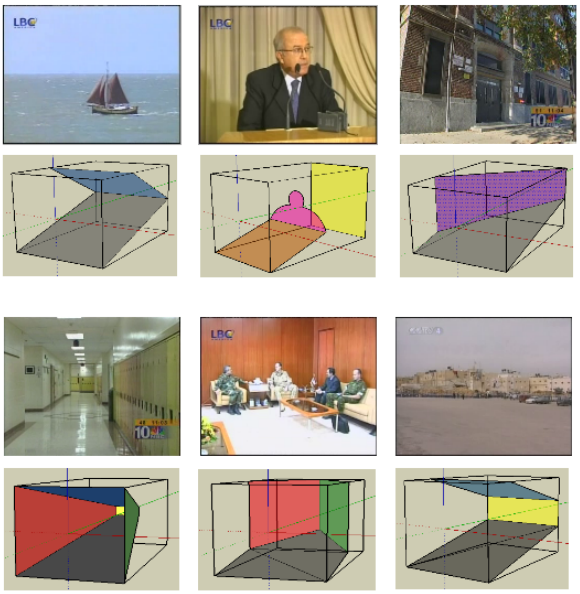

Figure 1. Example frames and their stage categories; top two rows, from left to right: *sky+ground*, *table+person+background*, *diagonal background*; bottom two rows: *box*, *corner*, *sky+background+ground*.

from a single arbitrary photograph.

Therefore, we do not aim for a precise reconstruction of the scene geometry. Accurate techniques have been designed for the reconstruction of the 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.

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, the absolute size of an object relative to the position in the scene, *etc*.

Recent methods for the geometry of scenes [9, 28, 20, 26] aim also at inferring the objects' geometry. There is a chicken-and-egg problem here: once the coarse geometry of the scene is known, one is able to deduce object sizes and use the information for object recognition, in a similar vein as Hoiem *et al*. [10]. However, learning the geometry may profit from recognizing familiar objects with known geometry such as faces, as exploited by Sudderth *et al*. [26].

We follow a different path. Inspired by the recent success of scene appearance classification [15, 6, 19, 30], into classes like indoor, outdoor, desert, beach and so on we consider classifying the stage type - that is, the rough geometry of the scene - on the basis of the regularities indicated above.

Scene classification methods capture the complex statistics of natural images by using bag-of-feature [22, 4] or texton approaches [13, 21, 31] to condense the scene appearance into codebooks, which is subsequently used to train a classifier from many examples of scene classes. The key ingredient here is the capturing of natural image statistics, as realized in the influential work of Torralba and Oliva [15, 28]. For real-world scenes, physical processes that shape natural structures are different at each observed scale [28]. These processes depend on object and material surfaces from which the visual world is built. Similar is true for man-made structures, which also differ due to functional constraints in relation to human size. Furthermore, the viewpoint with respect to horizon or vanishing point imposes constraints on image content. Image features capture these effects, and hence there is a relation between image statistics, scene structure and depth pattern.

That depth can also be derived from models of natural image statistics has already been shown by Torralba and Oliva [28]. However, whereas they propose the usage of mean depth information to facilitate scene categorization, we attempt to achieve the opposite.

In this paper, we aim at inferring geometry from single images. Implicitly, current methods for depth estimation from single images assume scene content to be classified, as they work for specific domain of indoor [26, 5], outdoor [9, 20], or have been specifically trained for a few of these categories [11]. We make this dependence explicit: we believe that the first step in providing depth information for a particular scene should be to classify the scene into one of the geometric stages (Figure 2). We build on the success of scene classification by learning a classifier to distinguish over various 3D stage categories. As the variety of stage types is much lower than the variety in pictorial scene content, we expect our method to be more successful than alternative methods directly based on the content.

In this paper we limit ourselves to the determination of the stage type. In the next phase, more precise depth estimation can be performed, by estimating the stage parameters from the data. Here, we will present stage classification results for the domain of TRECVID news videos [23].

The organization of the paper is as follows. In Section 2 we review related work and state our contribution. Section 3 outlines our approach to stage type classification. We present our results on news video data in Section 4. We wrap up with conclusions in Section 5.

2. Related work

Absolute depth from single images. Recent attempts to estimate absolute scene depth from single images use machine learning methods to directly map low-level features to image depths. Torralba and Oliva [28] use global image structure; based on the magnitude of the global Fourier transform and the local wavelet transforms, they obtain an estimate of the viewpoint and of average scene depth. Saxena *et al*. [20] also presented a method to learn absolute depth from single outdoor images based on low-level features, extracted at multiple scales, in a Markov Random Field (MRF) model. Delage *et al*. [5] derive an algorithm for reconstructing indoor scenes from a single image by learning the wall-ground boundaries using a Bayesian network.

For convincing visual 3D quality, and for many applications, including robot navigation, derivation of exact distances to elements in the scene may not be necessary as long as relative order of those elements is established. There exists a vast body of literature on recovering relative depth information. However, classical methods for relative depth estimation provide only local depth estimation and require high-quality images, as is the case for texture gradients [1], shape from shading (e.g. [12]), from edges and junctions

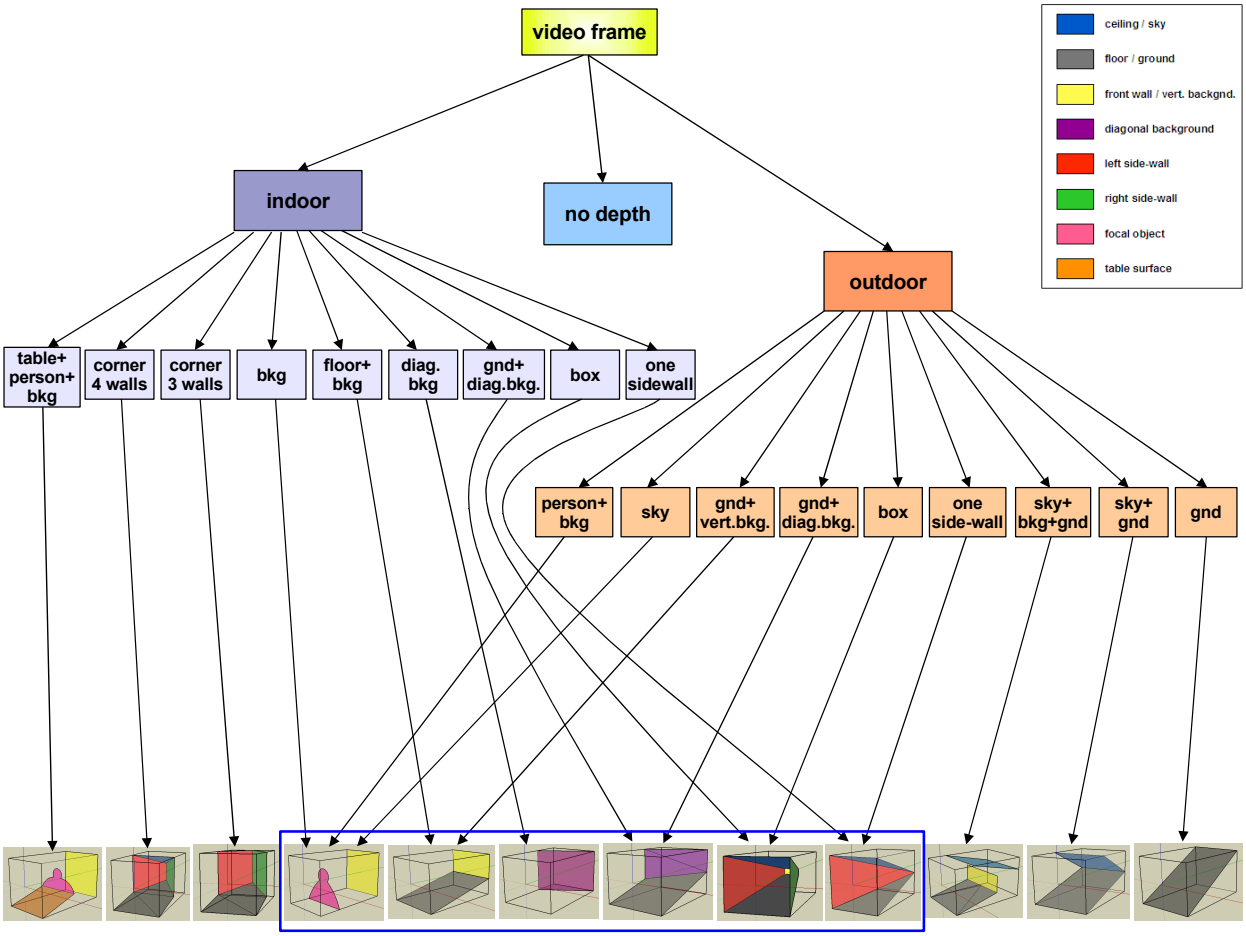

Figure 2. The original classification of TRECVID key-frames into stage types, inspired by Hoiem *et al*. [9] - the top-level categorization is into *indoor* and *outdoor* scenes, as well as those without a three-dimensional structure (i.e. *no depth*). However, the blue bounding box indicates a significant number of overlapping indoor and outdoor stages that share the same geometric model. Legend in the top right corner shows the labeling of individual surfaces inside stages.

[2], and from Fractal dimension [17] (see Palmer [16] for an overview).

Semantic scene categorization. Several researchers have constructed algorithms that can classify images into two semantic categories: indoor versus outdoor [27], city/suburb versus landscape [29], etc. They usually rely on particular discriminating features, for example that cities will have more vertical edge energy than flat landscapes. This last claim was used also by Oliva and Torralba [15], which propose a set of perceptual dimensions (i.e. naturalness, openness, roughness, expansion, ruggedness) to represent the dominant spatial structure of a scene. The same idea was used in [32] prior to learning the semantic scene context. A different approach, using a pre-defined codebook vocabulary, was used in [6] and [30] to label parts of an image by the best representative.

Geometric scene categorization. The spectral signature, used by Oliva and Torralba for scene categorization and estimation of mean absolute scene depth, has already proven useful in object recognition [14, 25]. However, scene classification approaches mentioned above suffer from two drawbacks. The first is that they model semantic scene categories. The potential number of such categories can be very large, and deriving high-level semantic information from images remains difficult and unreliable. The second drawback is that all these approaches work in the 2D image plane, without attempting to recover the 3D scene structure. To that regard, *geometric* image context has recently been used instead of semantic class modeling by Hoiem *et al*. [9, 8]. They model 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. Their model takes into account the observer's

viewpoint and the orientation of a physical object with relation to the physical scene.

Contribution of the paper. We draw inspiration from the work of Hoiem *et al*. [9] and attempt to derive 3D geometry of the scene, and recover depth information. However, instead of individual surfaces, we model geometric scene classes, relying on constraints imposed by both natural image statistics and viewpoint characteristics. We explicitly rely on categorization of the input image into stage types for proper estimation of scene geometry. 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 [28]. But where they propose to utilize mean absolute depth in order to facilitate scene categorization, we attempt to do the opposite, and propose to derive global depth profile based on stage types. They rely on natural image statistics as well for scene categorization [15], however we impose an additional constraint on global depth expectation since 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

As claimed in previous sections, we believe that the first step in providing depth information for a particular scene should be to classify the scene into one of the stages, each having a unique 3D geometry. Our stages thus refer to theatrical representations of physical scenes that provide specific context for objects, similar to cut&fold reconstructions of Hoiem *et al*. [9, 8]. Once the stage type is identified, the image can be aligned with its corresponding template, whereas individual objects can be placed in this 3D setup like cardboard figures.

3.1. Empirical study on TRECVID data

We relied on the structure present in real-world scenes in order to arrive at a limited number of geometric patterns. The structure of the visual world is imposed by three crucial constraints mentioned before: natural image statistics results in statistical regularities; viewpoint constraints (including the camera height typically $1.5-2m$) limit the possibilities with respect to perspective; and film rules ensure for the orthogonality of relevant lines and angles.

Our initial stages are shown in Figure 2: we have looked at thousands of TRECVID keyframes [23] and noted the frequency with which each specific category appears. The structure that we observed limited the vast number of surface combinations into 18 categories only, plus an additional *no depth* class corresponding to graphics (i.e. maps, charts, *etc*.). We retained only those stage types that corresponded to at least 5% of the observed video frames; this accounted for a large majority of all the data.

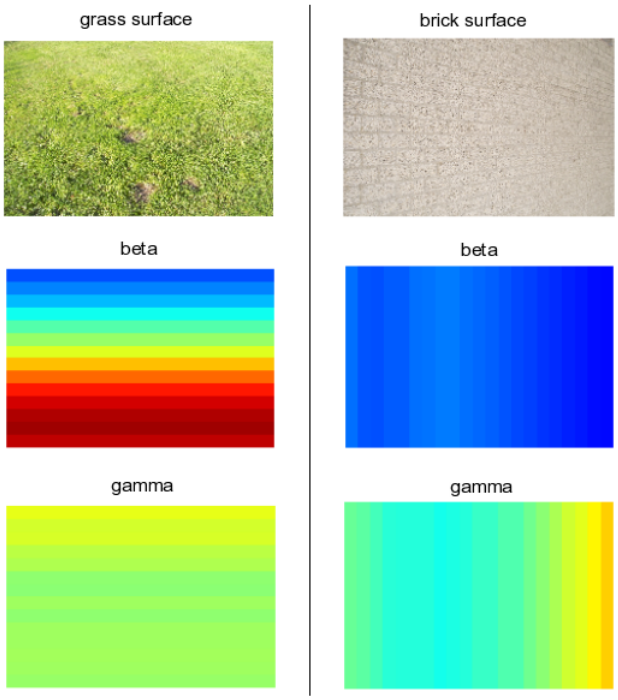

Figure 3. Natural image statistics-based Weibull distribution: parameter values as a function of depth for textures of grass (left column) and wall bricks (right column): β decreases from the point of fixation, whereas γ increases with depth. The color map has been hard-limited to a pre-defined range, such that results are comparable to each other.

When one models semantic scene classes, the natural top-level categorization is into indoor and outdoor images. In the beginning, we followed the same intuition and modeled outdoor stages as combinations of three basic layers (namely *sky, vertical background* and *ground*) and indoor stages as specific configurations inside a rectangular box. However, by observing many more frames and by representing the stages graphically, we have noticed that there exists a significant overlap in terms of geometry between these two top-level categories (see blue box in Figure 2). In other words, when geometrical classes are being modeled, the imposed structure leads to certain stage types being represented by the same 3D model, *regardless* of whether they belong to indoor or outdoor scenes. Therefore, we have concluded that higher levels of stage hierarchy should also be based on geometry.

3.2. Depth from stage types

Once we have successfully identified a small set of stage types, we turn our attention to descriptive visual features. As already mentioned, 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, with a single dominant structure observed, gradient histogram typically follows a decaying power-law distribution. When scene depth increases and more and more objects are added to the scene, 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 [7], whose parameters are indicative of local depth order and the direction of depth. This is shown in Figure 3 for two example surfaces. Spatial image statistics will conform to the Weibull distribution until the scene depth increases to the point that the observed samples become completely uncorrelated, resulting in a Gaussian histogram. Thus we capture natural image statistics by parameterizing edge histograms. We build on previous success of Weibull features in scene categorization [30] and generic concept detection [24] to classify scenes into stages and utilize this information for depth estimation.

Filtering. We use a Gaussian scale-space framework to extract features. Spatial scale is incorporated by convolving images with Gaussian derivative filters,

$$E\left(x,y,\sigma_i\right) = G\left(x,y,\sigma_i\right) * I\left(x,y\right) \tag{1}$$

where $G(x,y,\sigma_i)$ represents a Gaussian derivative filter in the x and y-direction, respectively, and I represents an intensity image.

Weibull distribution. From research on natural image statistics it is known that histograms of derivative filter responses can be represented by a simple distribution [18]. We follow [7] by exploiting the fact that histograms of gradient magnitude can be well modeled by an integrated Weibull distribution, also known as Generalized Laplacian,

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

The parameters μ, β and γ represent the center, width and shape (i.e. peakness) of the distribution, respectively, and x is an edge response of a derivative filter. Furthermore, $\Gamma(1/\gamma)$ denotes the complete Gamma function, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$. The shape parameter γ ranges from 0 to 2, however for general images it will often be within the interval [0.5,1]. For $\gamma = 2$ the Weibull distribution is equivalent to the normal distribution, and for $\gamma = 1$ it is a double exponential. For small values of γ, the distribution is close to the symmetric power-law.

Feature extraction. 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 μ, β and γ of the integral Weibull distribution. The μ parameter represents the mode of the distribution, whose position is influenced by uneven illumination. Therefore, in order to achieve illumination invariance, the values of μ are ignored.

By using Weibull parameters, an accurate and very compact parameterization of derivative histograms is obtained. The estimated distribution will fit well to histograms of the majority of all images, regardless of whether they are indoor or outdoor, artificial or man-made, *etc*.

Since our stages often contain oriented surfaces with continuously increasing depth, we need to perform local measurements and extract features from image regions. To that end, we have identified a standard grid for images, based on which we can perform the desired tasks. For each image, we extract grid-based 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 ($\sigma = 3$ pixels), resulting in β and γ parameters for each direction. Experiments are performed based on β and γ parameters together. Thus, for the total of 16 regions in our grid, we obtain feature vectors of 64 dimensions.

Classification strategy. The observations from our empirical study (Section 3.1) and Figure 2 led to a new, geometric hierarchy of stage types, in which only 15 geometric stages remain (including the symmetrical variants within certain classes). This is shown in Figure 4, according to which classification can be performed at an intermediate level (i.e. level of stage groups represented by Roman numerals) or at a lower level (i.e. individual stages - represented by Arabic numbers). Although the dataset is divided into 15 stage types, for reasons of clarity we decide to combine the results of symmetrical sub-stages, such that they are presented for 12 classes only. Figure 4 directly reflects our classification strategy.

For purposes of stage classification, 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'. The classification scheme is shown by a simplified block diagram in Figure 5.

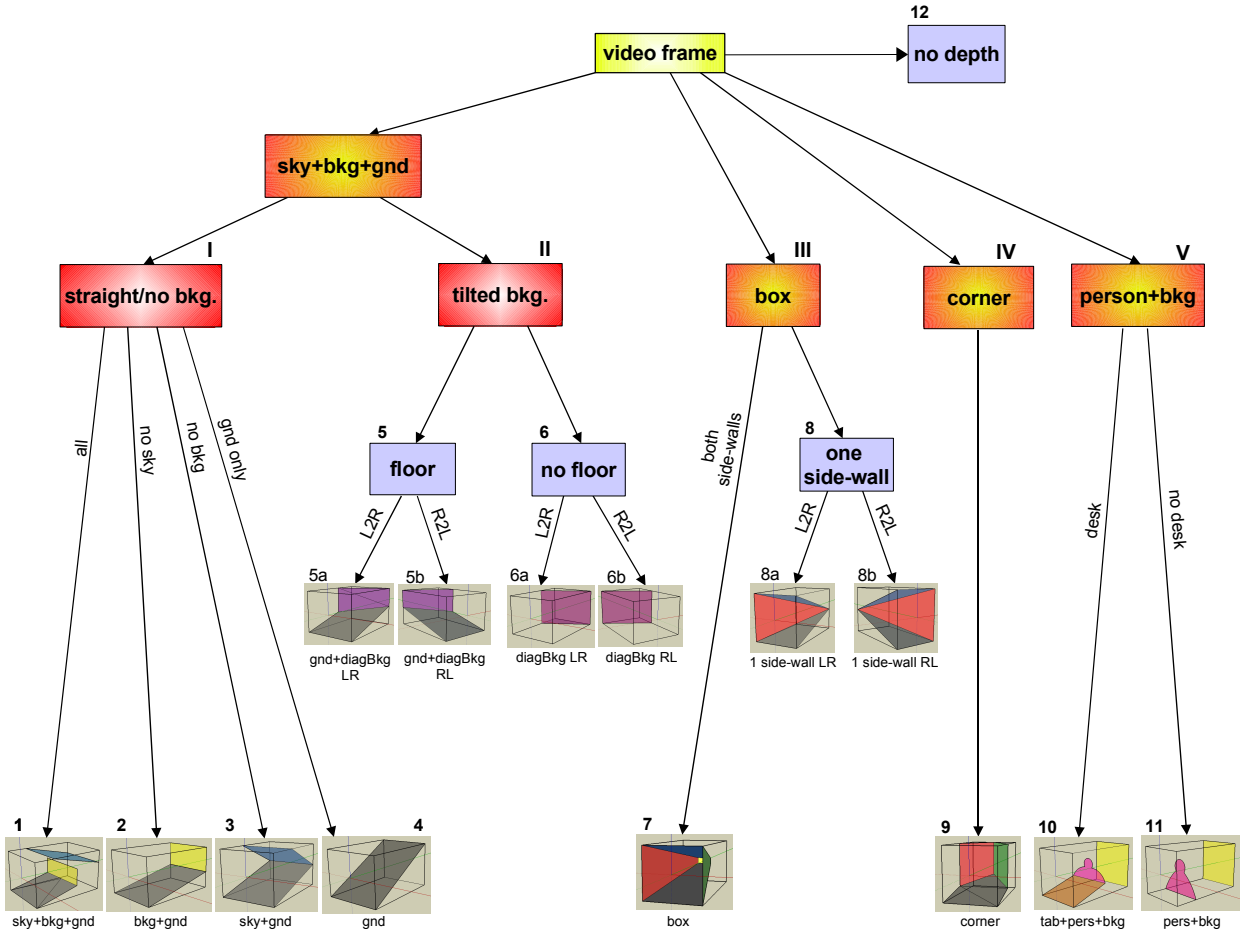

Figure 4. Hierarchical classification into stages: the classes of the intermediate hierarchy level are represented by Roman numerals I-V, whereas those of the lowest level are represented in Arabic numerals 1 through 12. Note that the symmetry of certain stages is represented by additional division into sub-stages.

4. Experiments

4.1. Experimental setup

For the evaluation of our stage classification algorithms, we have used the keyframes of the 2006 TRECVID video benchmark dataset[1]. The TRECVID video benchmark provides nearly 170 hours of news video in various languages (English: CNN, NBC, MSNBC; Chinese: CCTV4, NTDTV; Arabic: LBC).

In the initial result phase, we have annotated 1241 TRECVID keyframes into one of the 15 stage categories. Then for each category, samples were split before classification into two halves, one for training and another for testing purposes. From a large variety of supervised machine learning approaches, we have chosen the Support Vector Machine (SVM), which has proven to be a solid choice. We utilized the LIBSVM implementation [3] with radial basis functions as kernels.

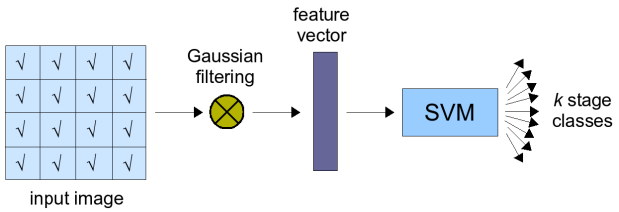

Figure 5. A simplified block-diagram of classification with Support Vector Machines: a generic stage detector is defined on the 4×4 region grid - features extracted from image regions are input into the SVM directly.

[1]NIST:TRECVID Video Retrieval Evaluation, 2001-2006. http://www-nlpir.nist.gov/projects/trecvid/

4.2. Results

Lower level of hierarchy - 12 classes. First we give results obtained with the 12 individual stages. Our stage classifier results are shown in Table 1, together with the relative occurrence (i.e. prior probability) of each stage type in the ground truth (as already mentioned, the results for symmetrical variants of the stages *ground+diag.background*, *diagonal background* and *1 side-wall* have been combined for clarity). The correct classification performance is given by the total number of correctly classified (true positives+true negatives) images divided by the total number of images. Our method performs very well for *sky+ground*, *gnd*, *table+person+background*, and *person+background* stages. Especially for the latter two classes, variety in image content is restricted for the domain of news videos, explaining the good performance. For less strictly defined stages, *e.g. ground+diag.background*, *corner* and *no depth*, performance is moderate due to the larger variability of scene content and exact scene configuration. For the classes *box*, *ground+background* and *1 side-wall*, performance is poor. Here, the drawback is the diversity of objects and amount of occlusions present in these categories. Hence, the statistics of the scene are lost in the object clutter.

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%

Table 1. Relative occurrence within the dataset and the percentage of correct classifications for the 12 stages at the lower level of hierarchy. For correct classifications, last row gives the average percentage over all classes.

Intermediate level of hierarchy - 5 classes When considering stage classification at a higher level in the hierarchy, more similar to the work by Hoiem *et al*. [10], the relative occurrences and results are given in Table 2. In this case, *straight/no background* and *person+background*, being the super-classes of well-performing stages from above, are again doing very well. On the other hand, *box* and *corner* are performing less good, due to the same reasons as before, namely the likelihood of object clutter and occlusion. Overall, a recognition performance of 40% is obtained.

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%

Table 2. Relative occurrence within the dataset and the percentage of correct classifications for the 5 stage groups at the intermediate level of hierarchy.

In conclusion, the results indicate that some simple stages (as well as their super-stages) can be detected very robustly. This is true for those classes which typically appear with small variations and are not likely to contain 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 samples, amount of variation within the class, significant amount of occlusion and object clutter, *etc*. Similar observations can be made with respect to the intermediate level of hierarchy with 5 stage groups. However, in all cases the performance is significantly better than the chance level, indicating the usefulness of the approach.

5. Conclusions

In this paper, we describe how the problem of depth information 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 having a unique depth pattern and providing a specific context for stage objects. Beside providing a background depth profile, this type of information about the scene significantly narrows down the possibilities with respect to individual objects' locations, scales and identities, and thus leads to more robust depth estimation.

Contrary to other scene classification approaches, we model geometric scene classes and thus account for the 3D relationships between objects and the scene. By relying on inherent structure of real-world images, resulting from natural image statistics and viewpoint constraints, we arrive at only 15 geometric stages for the news videos. We show that the proposed features are indeed indicative of depth information. Quantitative classification results are presented for the news video data of the TRECVID 2006 benchmark, yielding a baseline performance for stage type classification

in depth estimation. The results indicate that some simple stages, which typically do not appear with much variation and do not contain object clutter, can be detected robustly, with up to 60% success rate. Overall, classification performance for individual stages is around 28%, which may seem low. However, it should be considered that generic concept detection of video data in the NIST TRECVID benchmark reaches some 30% recognition rates, after many rounds of performance upgrades over the last 5 years.

In future work, we plan to utilize the hierarchy of stages to arrive at an individual stage after a two-step classification process. In other words, once the stage group for the image has been found, the individual stage is sought for only within the members of that group (i.e. within its "geometrical neighborhood"). Preliminary results show this strategy to be a very promising one.

It is important to note that in the presented work, we do not attempt to derive a precise depth map for the input image, but only to decide on the appropriate stage. However, the stage information constitutes a prior for 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.

References

[1] R. Bajcsy and L. Lieberman. Texture gradient as a depth cue. *Computer Graphics Image Processing*, 5:52–67, 1976.

[2] H. G. Barrow and J. M. Tenenbaum. Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence*, 17(1-3):75–116, 1981.

[3] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001.

[4] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

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

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

[7] J. Geusebroek and A. Smeulders. A six-stimulus theory for stochastic texture. *IJCV*, 62:7–16, 2005.

[8] D. Hoiem, A. A. Efros, and M. Hebert. Automatic photo pop-up. In *ACM SIGGRAPH*, 2005.

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

[10] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *CVPR*, 2006.

[11] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, October 2007.

[12] T. Kanade. Recovery of the three-dimensional shape of an object from a single view. *Artificial Intelligence*, 17(1-3):409–460, 1981.

[13] T. Leung and J. Malik. Recognizing surfaces using three-dimensional textons. *ICCV*, 1999.

[14] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: A graphical model relating features, objects and scenes. *NIPS*, 2003.

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

[16] S. E. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, 1999.

[17] A. P. Pentland. Fractal-based description of natural scenes. *IEEE PAMI*, 6:661–674, 1984.

[18] J. Portilla and E. P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *IJCV*, 40(1):49–70, 2000.

[19] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. *ICCV*, 1:883–890, 2005.

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

[21] C. Schmid. Constructing models for content-based image retrieval. *CVPR*, 2001.

[22] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. *ICCV*, 2003.

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

[24] C. G. M. Snoek, M. Worring, J. M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE PAMI*, 28, 2006.

[25] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. *ICCV*, II:1331–1338, 2005.

[26] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Depth from familiar objects: A hierarchical model for 3D scenes. *CVPR*, 2006.

[27] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *IEEE Int'l Workshop on Content-based Access of Image and Video Databases*, 1998.

[28] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE PAMI*, 24(9):1226–1238, Sep. 2002.

[29] A. Vailaya, A. Jain, and H. J. Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31(12):1921–1935, Dec. 1998.

[30] 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.

[31] M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. *ECCV*, 2002.

[32] J. Vogel and B. Schiele. Natural scene retrieval based on a semantic modeling step. In *Int'l Conf. on Image and Video Retrieval*, pages 207–215, 2004.