Stages as models of scene geometry

Vladimir Nedović, *Student Member, IEEE,* Arnold W.M. Smeulders, *Senior Member, IEEE,* and André Redert, *Member, IEEE,* Jan-Mark Geusebroek, *Member, IEEE*

Abstract—Reconstruction of 3D scene geometry is an important element for scene understanding, autonomous vehicle and robot navigation, visual inspection and 3D television. We propose accounting for the inherent structure of the visual world when trying to solve the scene reconstruction problem. Consequently, we identify scene categorization as the first step towards robust and efficient depth estimation from single images. We introduce 15 typical 3D scene geometries called *stages*, each with a unique depth profile, which roughly correspond to a large majority of all images. Stage information serves as the first approximation of global depth, narrowing down the search space in depth estimation and object localization. We propose different sets of low-level features for depth estimation, and perform stage classification on two diverse datasets of television broadcasts. Classification results demonstrate that stages can be efficiently learned from low-dimensional image representations.

Index Terms—scene geometry, scene structure, depth estimation, scene categorization, stages

1 Introduction

Visual perception is the process of inferring world structure from image structure. Although the projection of the physical 3D scene onto the image plane carries ambiguities as to which physical configuration gave rise to the depiction, human observers can effortlessly derive an impression of scene depth from a single image. This is because the world around us behaves regularly, and because structural regularities are directly reflected in the 2D image of some world scene [1]. The aim of this paper is to identify these regularities in 2D images and exploit them for the purpose of scene geometry reconstruction of generic image content.

Consider examples in Figure 1. In the regular world that we live in, certain scene configurations tend to appear significantly more often than others. There are rough classes of scene geometries, which we call *stages*, that include a straight background (like a curtain, a wall, the façade of a building, a remote mountain range), or other ones which show walls at three sides of the picture (a corridor, a tunnel, a narrow street). If television broadcasts are considered, there is also a specific stage for anchor-type images, corresponding to news-reader sequences, interviews, talk-shows and press-conferences. Figure 1 shows a few prototypes together with their stage models.

We have arrived at the notion of a stage from the observation that objects of the world act in relatively stable, recurring geometrical environments. The objects come with almost infinite variation in appearance, as well as geometry. Scenes, on the other hand, show a much more regular pattern. One is quickly led to conclude that a vast majority of images depicts the scene with only a few different geometry types. This is not surprising

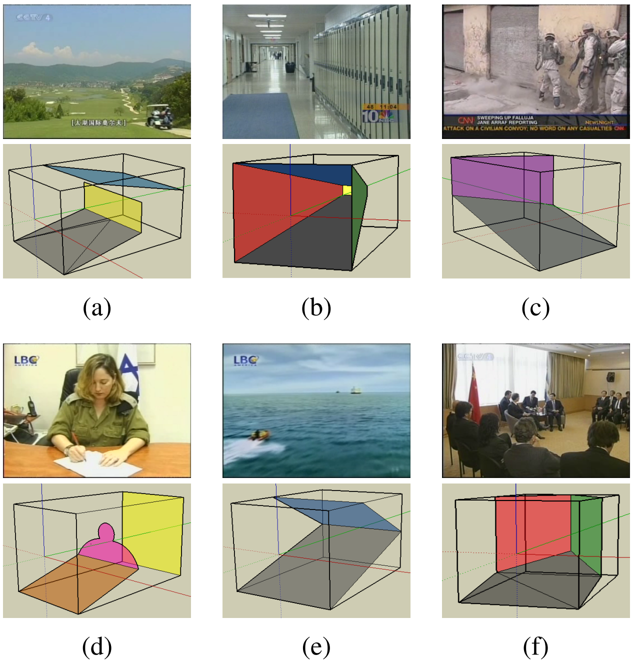


Fig. 1: Example frames and the corresponding stage models, containing at most five planes. Each image carries lots of information about global scene depth. Note in (a) that colors and contrasts fade towards the horizon. Image in (b) offers plenty of tilted perspective lines from which depth directionality can be determined. Example in (c) demonstrates that, although the objects might move around considerably, scene geometry is not likely to change substantially. In (d), lots of motion is not expected for the large foreground object either, making it an integral part of the scene. In the image shown in (e), the texture grain continuously decreases in size towards the horizon, for both water and sky surfaces. For complex scene configurations, such as shown in (f), a combination of texture and perspective information is needed to estimate depth.

- *Vladimir Nedović, Arnold W.M. Smeulders and Jan-Mark Geusebroek are with Intelligent Systems Lab Amsterdam (ISLA), University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands.*
E-mail: {vnedovic,smeulders,mark}@science.uva.nl
- *André Redert is with Philips Research Laboratories Eindhoven, High Tech Campus 36, 5656 AE Eindhoven, The Netherlands.*
E-mail: andre.redert@philips.com

if one considers only a subset of given constraints. Namely, straight image lines always converge at vanishing points and the horizon; walls are almost always perpendicular to the ground surface; the camera is almost always at approximately two meters height, etc. The important conclusion is that whereas a precise geometry may be requested for the object, for many applications it suffices to build a rough model for the geometry of the scene. In this paper, we aim to discover whether stages as models of scene geometry can be derived from a single image.

There are many advantages of knowing just the stage type. Apart from identification of the scene configuration, the stage may reveal to the observer information about the semantic context of the scene, the identities of scene elements, as well as relative depth order. Recognition of the stage thus serves as the entry point of a detailed depth analysis, object localization and object recognition.

We do not aim for a precise reconstruction of scene geometry. Accurate techniques have been designed for the reconstruction of object geometry via shape from shading [2], shape from motion (e.g. [3]) or shape from stereo (e.g. [4]), when these options are available. The scene geometry, in contrast, is the stage on which objects of the picture act, hence limited accuracy frequently suffices. Here we consider stages as rough models of the scene, with the objects ignored. Motivated by the recent success of scene appearance classification into classes like indoor, outdoor, desert, beach, and so on, we consider classifying the stage type on the basis of regularities indicated above. Given systematic changes in image perspective, texture and colors over depth, we focus on a small set of features carrying the depth information in general scenes.

1.1 Contribution of the paper

We draw inspiration from the work of Hoiem et al. [5], [6], attempting to derive piecewise planar 3D geometry of the scene and corresponding depth information. They assume the presence of certain surfaces in the image, such as sky and ground, which limits the applicability of their approach. Instead, we attempt to develop a more general method for scene reconstruction and follow a different computational path. Whereas they learn the geometry of individual surfaces, we take a holistic approach by modeling scene classes, relying on constraints imposed by image structure, texture and colors.

Our work on depth estimation is also inspired by Torralba and Oliva [7], who have utilized models of natural image statistics to derive depth. But where they propose to use mean absolute depth to facilitate scene categorization, we attempt to do the opposite, and derive global depth profiles based on stage types. Beside exploiting image statistics, we impose additional constraints, which greatly reduce the number of categories that we need to model.

In this paper we limit ourselves to the determination of the stage type, extending the work from [8]. In the next phase, more precise depth estimation can be performed, or stage information can be used for object localization. We present a variety of experiments for the domain of 2006 TRECVID news videos [9], and extend this evaluation with a test on an independent dataset of our own 2007 television recordings.

2 Related work

2.1 Depth estimation from a single image

Recent methods for estimating scene geometry [10], [7], [11], [12] aim at inferring the geometry of objects as well. This approach suffers from a chicken-and-egg problem: 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. [13]. However, learning scene geometry may profit from recognizing familiar objects with known 3D shapes, as has been shown for office objects by Sudderth et al. [12].

Attempts to estimate absolute scene depth from single images use machine learning methods to directly map low-level features to image distances. Torralba and Oliva use local and global image structure to derive average scene depth [7]; Saxena et al. learn absolute depth from features at multiple scales [11], [14]; and Delage et al. [15] reconstruct indoor scenes by learning the wall-ground boundaries.

However, for many applications, derivation of exact distances to elements in the scene may not be necessary, as long as relative order of those elements is established [16]. There exists a vast body of literature on recovering *relative* depth information. However, classical methods in this field provide only local depth estimation and require high-quality images, as is the case for texture gradients [17], shape from shading (e.g. [18]), from edges and junctions [19], and from fractal dimension [20] (see Palmer [21] for an overview). Recently, Hoiem et al. [6] proposed a method for determining relative depth order of prominent scene surfaces. However, they build on previous work on scene reconstruction [5], which assumes the presence of sky, ground and vertical surfaces, and is thus constrained to only some of the typical scene configurations.

2.2 Scene categorization

Several researchers have constructed algorithms to classify images into two semantic categories: indoor versus outdoor [22], city/suburb versus landscape [23], etc. They usually rely on particular discriminating features, for example that cities have more vertical edge energy than flat landscapes. The key ingredient here is the capturing of natural image statistics, as realized in the influential work of Oliva and Torralba [24], which represent the spatial structure of a scene using a set of perceptual dimensions. A different approach, with a predefined codebook vocabulary of visual words, was used in [25], [26], [27] and [28] to label parts of an image by the best representative. This was subsequently extended in [29] with a spatial pyramid for multi-scale features.

Scene classification approaches mentioned above have 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 semantics 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 proposed instead of semantic class modeling in [5]. They define classes of image surfaces

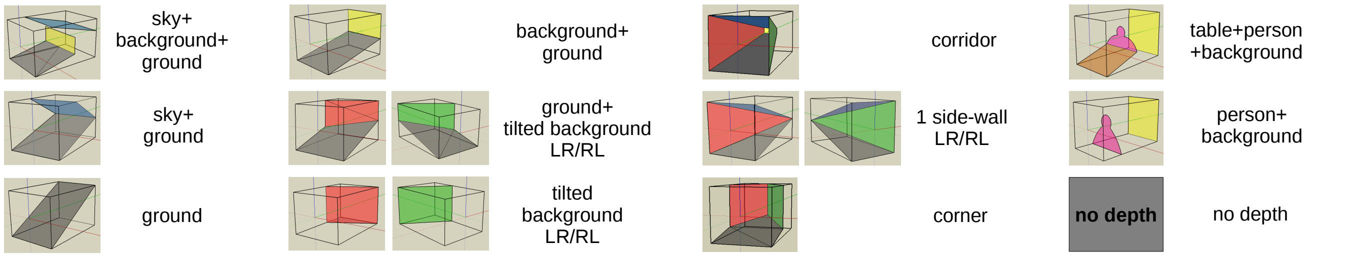


Fig. 2: The structure of visual world leads to only 15 typical scene geometries - *stages*. We represent their depth profiles by simple, piecewise-planar models that serve as the first approximation to background scene depth.

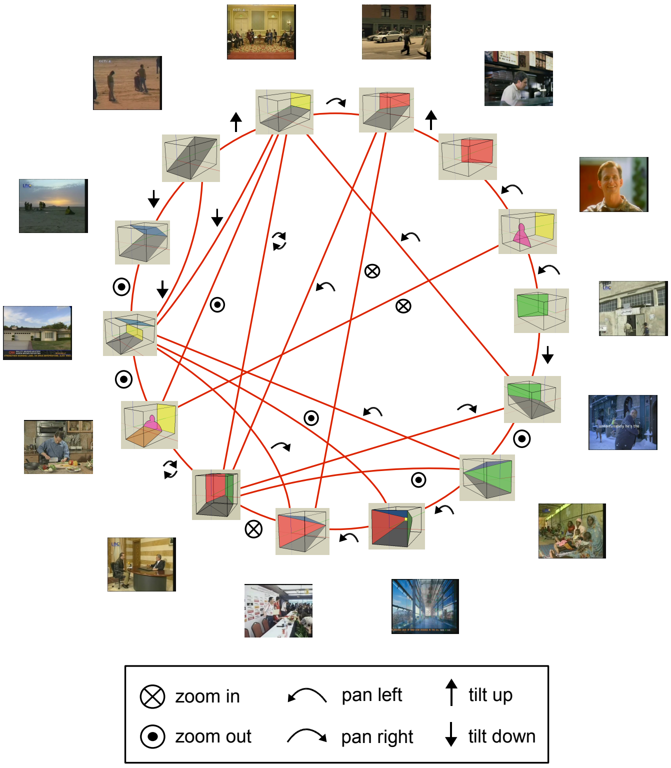


Fig. 3: Stage transition graph in terms of elementary camera movements, demonstrating the stage continuum concept. symbols next to the connections denote clockwise (or, for connections inside the circle, left-to-right) transitions. Real images included next to the stage models represent typical examples.

and learn their orientations; the subsequent combination of the surfaces leads to the reconstructed 3D scene model. In contrast, we believe that learning the geometry from scene, as opposed to surface classes, is a much simpler problem than image segmentation and subsequent reconstruction.

Implicitly, current methods for depth estimation from single images assume scene content to be already classified, as they work for the specific domain of indoors [12], [15] or outdoors [10], [11], or have been specifically trained for one of these categories [5]. We make this dependence explicit: we believe that the first step in providing depth information for a particular scene should be to classify this scene into one of the geometric types - stages.

3 Depth from Stage Types

3.1 The structure of visual data

We rely on the structure of visual images in order to arrive at a limited number of geometric scene types. The structure of scenes is a consequence of three crucial phenomena. From literature on natural image statistics it is well known that 2D images exhibit statistical regularities [30]. However, these findings have so far been primarily put in the perspective of efficient image coding [31], and only recently similar ideas have been considered in the context of depth estimation. Torralba and Oliva have consequently shown that an estimate of average scene depth can be derived from natural image statistics-based features [7]. Furthermore, Yang and Purves have offered a statistical explanation for depth perception, which accounts for the properties of scene geometry relative to the observer [16]. In addition to image statistics, however, there are other factors which give rise to image regularities. In [1], the authors note certain scene configurations that occur much more frequently than others. They call such configurations "modal", pointing that they are characterized by orthogonality between lines and planes. Due to gravity, as well as features of man-made structures, such scene configurations are the most prominent in natural images. They are additionally emphasized by the composition and filming rules, whose conventions ensure the proper order of image surfaces (e.g. that sky is on top and ground is at the bottom). Finally, viewpoint constraints limit to a large degree the possibilities with respect to perspective. Relatively small range of vertical camera tilt, together with height typically at $1.5-2m$, determine the location of the horizon and the vanishing point(s), as well as the scales and positions of most viewed objects.

3.2 The stages

In order to identify prominent scene configurations, we have noted the frequency of specific surface combinations in several thousand TRECVID keyframes [9]. Since the emphasis has been solely on geometry, there emerged a significant overlap between indoor and outdoor configurations, avoiding the distinction between these two high-level categories [8]. In addition to the inherent structure of visual data, this limited the number of possible surface combinations to only 15 geometric categories, shown in Figure 2. The category set includes a *'no depth'* class, corresponding mostly to graphics frames (i.e. maps, charts, etc.). We have retained configurations that

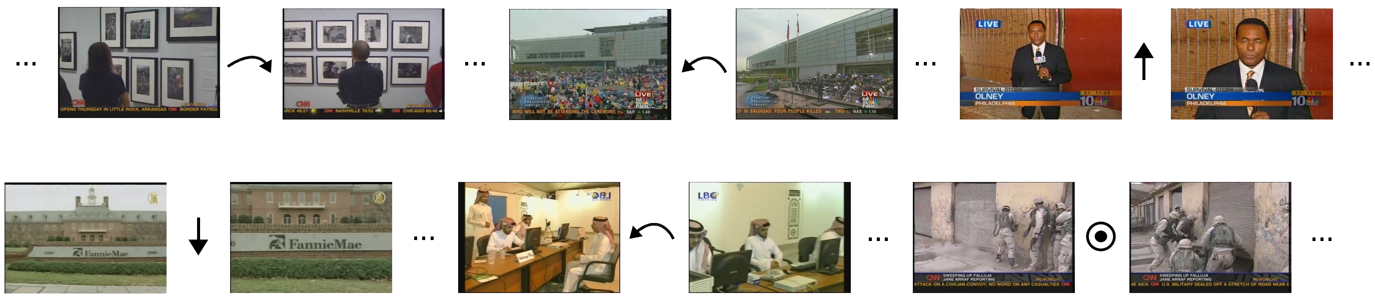


Fig. 4: Example sequences with frames undergoing stage transitions as formulated by Figure 3.

corresponded to at least 5% of the examined video frames; this accounted for a large majority of all the data.

An important remaining question is whether a small number of identified stages is sufficient to represent actual configurations appearing in images. To that regard, stages can be considered as samples in a continuous space defined by camera parameters. In such a space, stage transitions are also defined, in terms of basic camera operations such as pan, zoom, and tilt. If our models are indeed representative samples of the continuum, then the outcome of any camera movement from a given stage would not result in any scene configuration, but again in a known stage model. More formally, for input stage s_i, we have

$$f(s_i, \omega) = s_j, \qquad \omega \in \{tilt, pan, zoom\} \tag{1}$$

where it is possible to have $i = j$. For small values of the parameters, the given stage remains of the same type, but when a certain value is exceeded, it transitions into another category. Conversely, every stage should be reachable from some other stage by means of a single camera operation (that induces a sufficient change in the parameters). Figure 3 illustrates that these conditions are indeed fulfilled - it presents a transition graph with possible outcomes for many stage-parameter pairs. Although six transitions are in principle defined for every given stage, for reasons of clarity only certain important links are shown.

3.3 Visual Features

To derive depth information from images, we use four sets of descriptive visual features. For the reference set, we use the subset of the geometric context feature set, proposed in the state-of-the-art system of Hoiem et al. [10]. The second set includes texture gradient features - natural image statistics-based Weibull parameters. Representative atmospheric scattering features comprise the third set, accounting for changes in scene colors due to increased depth. Finally, the fourth set of features encodes information related to tilted perspective lines, via anisotropic Gaussian filtering.

Scene classification in general profits from local information. To that end, we define a 4×4 grid of image regions, each spanning $\frac{w}{4} \times \frac{h}{4}$ pixels, where w and h denote image width and height, respectively. For each image, we extract and concatenate the 16 local region measurements into a single feature vector used for classification.

Feature Descriptions	# ft.	used
Color	**16**	**Yes**
C1. RGB values: mean	3	
C2. HSV values: C1 in HSV space	3	
C3. Hue: histogram (5 bins) and entropy	6	
C4. Saturation: histogram (3 bins) and entropy	4	
Texture	**15**	**Yes**
T1. DOOG filters: mean abs. response of 12 filters	12	
T2. DOOG stats: mean of variables in T1	1	
T3. DOOG stats: argmax of variables in T1	1	
T4. DOOG stats: (max-median) of variables in T1	1	
Location and Shape	**12**	**No**
3D Geometry	**31**	
G1. Long Lines: total number in region	1	Yes
G2. Long Lines: % of nearly parallel pairs of lines	1	Yes
G3. Line Intscn: hist. over 8 orientations, entropy	9	Yes
G4. Line Intscn: % right of center	1	Yes
G5. Line Intscn: % above center	1	Yes
G6. Line Intscn: % far from center at 8 orient.	8	Yes
G7. Line Intscn: % very far from center at 8 orient.	8	Yes
G8. Texture gradient: x and y "edginess" center	2	No

TABLE 1: Geometric context feature set [10]. The second column shows feature dimensionality per image region, whereas the last one indicates whether or not some feature is used in our experiments.

3.3.1 *Geometric context features*

As a reference feature set, we have implemented the geometric context features used in [10]. These are summarized in Table 1. Contrary to Hoiem et al., who compute the parameters from segmented super-pixels, we derive them at the level of regions. Therefore, 'Location and Shape' parameters are not used, since all would result in fixed values. However, these parameters are implicit in our feature extraction and classification setup, and are thus superfluous.

Without the 'Location and Shape' subset, there are 60 features per image region, resulting in a 960-dimensional vector for the whole image.

3.3.2 *Texture gradient features*

There exists a direct relation between image statistics, scene structure and depth pattern [7]. When scene depth is small, larger surfaces will be coarse, showing smaller 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 objects are added to the scene, the texture of the image will be fragmented

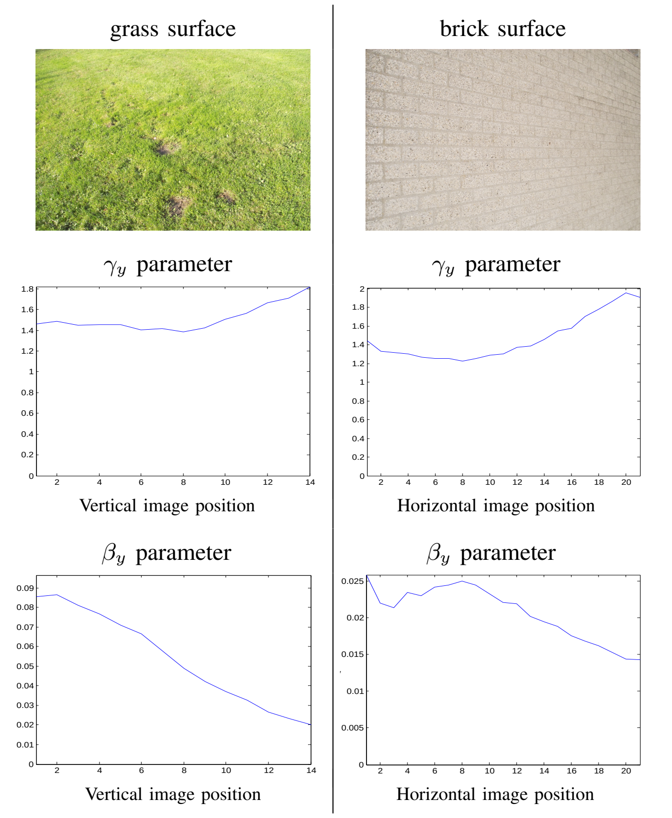


Fig. 5: Natural image statistics-based texture gradient features. Weibull parameter values (y-gradient) as a function of depth for textures of grass (left column) and wall bricks (right column).

into various patches, each associated with a different power-law. The integration over various power-laws results in a Weibull distribution [32], whose parameters are indicative of local depth order and the direction of depth. Spatial image statistics will conform to Weibull distribution until the scene depth increases to the point that the observed samples become completely uncorrelated, resulting in a Gaussian histogram.

Thus we follow [32] by exploiting the fact that histograms of gradient magnitude can be well modeled by an integrated Weibull distribution,

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

The parameters μ, β and γ represent the center, width and peakness of the distribution, respectively, whereas r is an edge response of a derivative filter (Γ denotes the Gamma function).

Using Gaussian derivative filters, we extract texture information that is subsequently summarized in histograms. We use a maximum likelihood estimator (MLE) to derive 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, we ignore μ to achieve illumination invariance.

Integral Weibull distribution is fitted to histograms of intensity filter responses in x and y directions (σ set to 3 pixels), resulting in β and γ parameters for each direction. Figure 5 demonstrates the change in Weibull parameters over depth, for two example surfaces. Histograms of y-gradient were computed from small squared regions, and individual Weibull parameters averaged along the direction perpendicular to change in depth. For both surfaces, γ increases with depth, whereas β decreases from the point of fixation.

Thus we capture natural image statistics by parameterizing gradient histograms. By using Weibull parameters, an accurate and very compact representation of the histograms is obtained. With only 4 features per region, there are 64 elements in the feature vector for the whole image. We build on previous successes of these features in scene categorization [28] and generic concept detection [33] to classify scenes into stages.

Feature type	# feat.
Texture gradients	**4**
- Integral Weibull in $\{x, y\}$-direction: β	2
- Integral Weibull in $\{x, y\}$-direction: γ	2

TABLE 2: Summary of our texture gradient features.

3.3.3 *Atmospheric scattering features*

The key properties of light, such as its intensity and color, are affected as the light travels through the atmosphere [34]. As a consequence of this *atmospheric scattering*, there exists a simple, non-linear relationship between intensity of light in an image and the physical distances between objects and a camera. More specifically, systematic changes occur in the contrast and color of objects when they are viewed from a distance [21]. When an open landscape is observed, for example, scene elements appear progressively “fuzzier”, and their colors lighter and more “washed out”, as the focus of attention shifts from the foreground toward the horizon.

Therefore, properties of the light source and colors of scene elements can be used as depth cues. This has already been shown in [35], where absolute depth for a small set of outdoor images is derived using a model of light scattering; and in [34], where depth segmentation is obtained from multiple images of the same scene, taken under different weather conditions. Since Weibull parameter β already encodes image contrast to some extent, we use other properties of light to represent the degree of atmospheric scattering. To that end, color saturation is utilized as an indicator of color purity, whereas an estimation of illumination color encodes the properties of the light source.

Color saturation reflects the amount of energy at color's dominant wavelength, relative to the amount of white light. In RGB color space, it corresponds to the distance outward from the intensity axis to the position of the color at each point. In our implementation, saturation is defined as

$$S = \frac{max\,(R,G,B) - min\,(R,G,B)}{max\,(R,G,B)} \tag{3}$$

For estimation of the illumination color, we use a well-known Grey-World algorithm, which assumes that the average reflectance in the image is achromatic. From each region, we extract distribution parameters mean (μ) and variance (σ^2) of the saturation component, as well as three coefficients corresponding to illumination color.

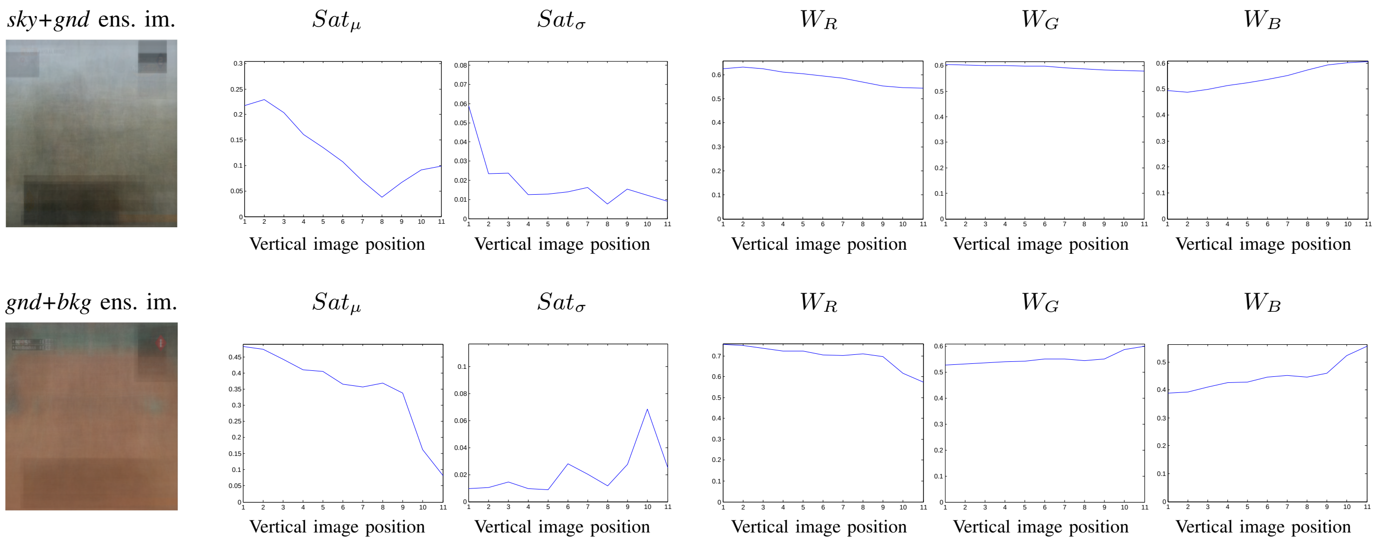


Fig. 6: Atmospheric scattering features as a function of depth, computed from mean class images. Color saturation mean (Sat_μ) and standard deviation (Sat_σ) parameters, and color correction coefficients for red (W_R), green (W_G) and blue (W_B) channels were extracted from 100 × 100 square regions of ensemble images, then averaged along the x-direction.

The effect of depth changes on these parameters is shown in Figure 6. For the two stage categories whose depth profiles show a continuous increase in depth from bottom to top of the image (namely *sky+ground* and *ground+background* types, containing 81 and 332 members, respectively), we have computed the mean image of all the samples. After some pre-processing of each image (described in detail in Section 3.4), the ensemble was obtained as the point-by-point average over all category members.

As shown in the figure, saturation clearly decreases with depth, with the variation in values depending on pre-processing. Color correction coefficients show interesting behavior. The coefficient for the blue channel increases with depth in both cases. This is expectable in images containing sky (such as those in the example in the top row), however it is the case also for those which have a straight vertical background of any type (i.e. the example in the bottom row). The red channel coefficient shows the opposite behavior, whereas the one for the green channel remains relatively stable. In fact, the three coefficients are dependable on each other whenever the illumination intensity ($I = R+G+B$) is evenly distributed over the whole scene. The assumption about constant illumination can safely be made for images without much shadows and shading.

The total number of features in the atmospheric scattering set is 80. This feature set contains a simple and compact representation of color properties which are altered with changes in depth.

Feature type	# feat.
Atmospheric scattering	**5**
- Color saturation: μ and σ^2	2
- Illumination color: $\{R, G, B\}$ coefficients	3

TABLE 3: Summary of our atmospheric scattering features.

3.3.4 Perspective line features

Some of the most useful information for depth estimation are structures which reveal distortions due to perspective projection. Straight lines, parallel in the observed scene but tilted in the image, and ultimately converging at vanishing points and the horizon, all contribute to determination of depth directionality.

The detection of directional structures is often unreliable when disturbing influences (such as noise, shadows and shading, blending object borders, etc.) have to be ignored. In these cases, it is desirable to have a detection method which ignores the distorting data aside the line, while accumulating evidence of the line data along its orientation. This implies a sampling of orientations by anisotropic filtering, suited for robust measurement of the differential image structure.

Geusebroek et al. show the decomposition of the anisotropic Gaussian filter in [36]. A filter at arbitrary orientation is decomposed into a sequence of two Gaussian one-dimensional filters in non-orthogonal directions:

$$g_\theta(x, y; \sigma_u, \sigma_v, \theta) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left\{-\frac{1}{2}\frac{x^2}{\sigma_x^2}\right\} * \frac{1}{\sqrt{2\pi}\sigma_\varphi} \exp\left\{-\frac{1}{2}\frac{t^2}{\sigma_\varphi^2}\right\} \quad (4)$$

Thus the sequence of two 1D Gaussian convolutions yields the anisotropic smoothed image. Since texture gradient parameters of Section 3.3.2 are extracted for x and y directions, we use the anisotropic filter at 4 tilted angles, namely for $\theta = \{30°, 60°, 120°, 150°\}$. Using MLE estimator, we parameterize the resulting gradient distributions by fitting a Weibull shape to each (see Section 3.3.2). This way, a compact descriptor of $4 \times 2 \times 16 = 128$ features is obtained.

Figure 7 shows how our perspective line features change with an increase in surface depth. As with the above texture gradient features, γ increases with depth, whereas β decreases from the fixation point.

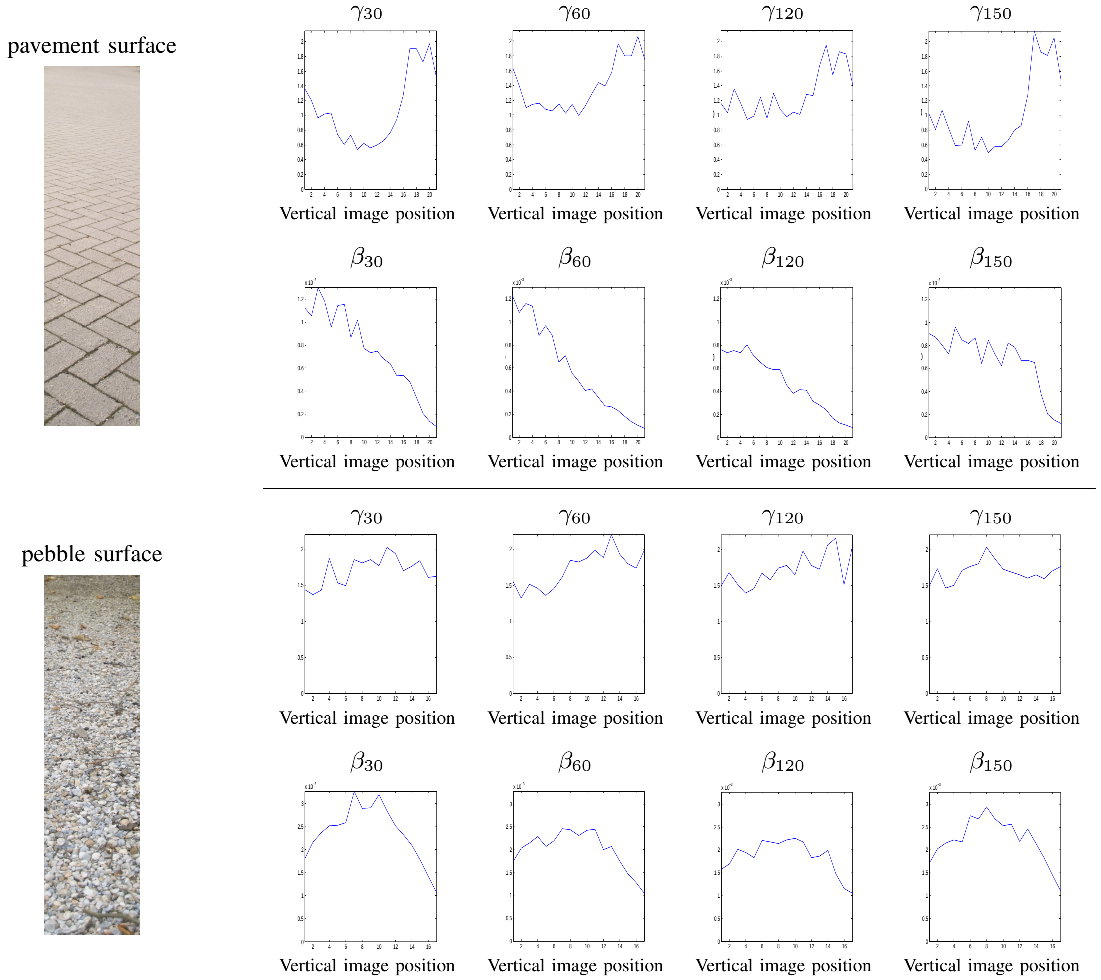


Fig. 7: Perspective line features, derived via anisotropic Gaussian filtering at $30°, 60°, 120°$ and $150°$. Weibull distribution parameters as a function of depth for pavement (top two rows) and leaves surfaces (bottom two rows). The parameters were extracted from 100×100 image regions, then averaged along x direction.

Feature type	# feat.
Perspective lines	**8**
- β at orientation $\{30°, 60°, 120°, 150°\}$	4
- γ at orientation $\{30°, 60°, 120°, 150°\}$	4

TABLE 4: Summary of our perspective line features.

3.4 Pre-processing steps

For our experiments, we use datasets which contain diverse television recordings from various international channels. In these datasets, image data can be distorted in many ways; beside compression artifacts that are common in other datasets, our frames may contain black bars around the image, as well as superimposed channel logos and subtitles. All this implies the necessity of pre-processing, so that only relevant pictorial data is extracted and used in further analysis.

Before feature extraction can be performed, uninformative parts of the image are removed in several steps, in a completely automatic manner. Dark borders (i.e. within 8% of black) are first removed around the whole image. Then upper image corners are checked for the presence of channel logos, whereas the bottom of the image is subjected to subtitle detection. In both cases, the detection is performed in gradient domain, where sharp edges for superimposed graphics are more easily revealed. Namely, if the gradient histogram of the relevant region is shifted towards very strong edges, the graphics are likely to be present, and their contribution to the image is removed. After the process, the original pictorial information remains of the same size as before (as our approach depends on the configuration of structures within the frame).

4 Experiments

4.1 The data

For the evaluation of our stage classification algorithms, we have used the keyframes of the 2006 TRECVID video benchmark dataset [9]. This benchmark provides nearly 170 hours of news channel videos in various languages (English: CNN, NBC, MSNBC; Chinese: CCTV4, NTDTV; Arabic: LBC). For purposes of classification, we have annotated a subset of 1315 TRECVID keyframes into one of the 15 stage categories. Beside these, there were 23 examples of other, less prominent surface configurations, and 76 examples of cases difficult to annotate. The difficult cases are usually images taken under low visibility conditions, or depictions of crowds, explosions with smoke and fire, intricate shapes, and objects in extreme close-up views. Figure 8 shows some examples.

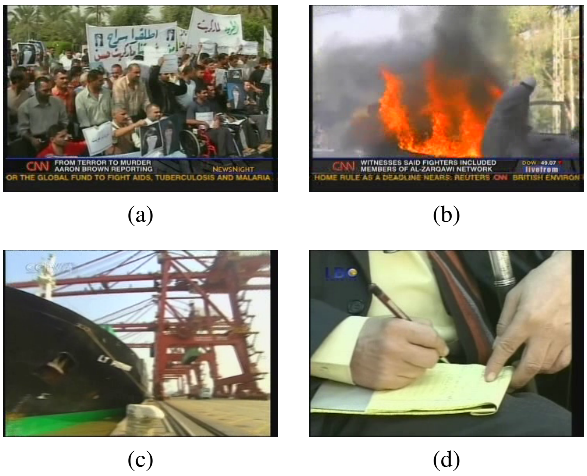

(a) (b) (c) (d)

Fig. 8: Examples of difficult cases for stage annotation. (a) Crowds. (b) Smoke and fire. (c) Intricate shapes. (d) Extreme close up view.

4.2 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' [37].

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 [38] implementation, with radial basis functions as kernels.

4.3 Description of experiments

4.3.1 Experiment I - Geometric context features

In the first experiment, we have used the 960-dimensional geometric context feature set. Although support vector machines are known to be robust with large vector spaces, they could not handle this highly-dimensional set very well, labeling all images by the same category. The SVM results are therefore not shown here. Instead, the numbers reported below have been obtained with an AdaBoost [39] algorithm, another linear classifier that proved reliable in practice.

4.3.2 Experiment II - Texture gradient features

For the second experiment, texture gradient features described in Section 3.3.2 are being used. A 64-dimensional feature vector containing Weibull parameters has been directly input into an SVM. Note that this experiment is related to the one reported in [8], except that in this case the frames are subjected to the chain of pre-processing steps described in Section 3.4.

4.3.3 Experiment III - Atmospheric scattering features

Here, we have used the atmospheric scattering features of Section 3.3.3. Information about color saturation and illumination color was used as an 80-dimensional feature vector in SVM.

4.3.4 Experiment IV - Anisotropic Gaussian features

For the fourth experiment, we use the perspective line features described in Section 3.3.4. The gradient distribution parameters corresponding to oriented structures at 4 angles are fused into a 128-dimensional feature vector.

4.3.5 Experiments V and VI - combined feature sets

In experiments V and VI, we test certain combinations of individual feature sets, which are concatenated together for that purpose. That way, the 'Texture+atm.scatter' combination results in a feature vector of 144 dimensions, whereas 'Atm.scatter+perspective' produces a 208-element vector. As texture gradient and perspective features are encoding similar information, their combination did not produce any improvement in the results, and is therefore omitted. Similarly, texture gradient information did not add value to the feature set containing all the proposed individual features together. The same was the case when 'Geometric context' features were included in any of the combinations.

4.4 Results

In this section, we present stage classification results, where we have combined the performance figures of symmetrical variants (i.e. *1 side-wall, tilted background* and *ground+tilted background* types), such that numbers are given for 12 classes only. Grouped symmetrical variants are represented with a single stage model, in which violet color indicates the tilted vertical plane both with left-to-right, as well as right-to-left increase in depth.

Instead of the usual percentage of all correctly classified samples, performance is given by per-class recognition rates. This manner of reporting results is much more suitable for problems involving uneven prior class distributions. In each experiment, five runs are performed, each time with a different random drawing of samples - one half for training, and the other for testing purposes. The results reported represent the mean average recognition rates over the five runs.

The results of all the experiments are summarized in Figure 9, which provides the comparison of different feature sets. In addition to classification performance of different features, the prior probability of each class is shown, which can be considered as the performance baseline. The confusion matrix for the best performing feature set is also shown in Figure 10.

From Figure 9, we see several trends for classification into 12 individual stages. For some classes, the performance is above 25% in most experiments, and above 35% with the best-performing feature set; this is the case for 8 out of 12 stages. On the other hand, the recognition of 4 remaining types, namely *corridor, tilted background, ground+background* and *1 side-wall*, is relatively poor. This is due to the larger variability of scene configuration, the diversity of objects and amount of occlusions present in these categories. Most of the feature sets result in recognition rates that are significantly higher than prior probabilities for each class. The gap between the baseline and the best result is biggest for stages *sky+ground* (60.3% difference), *person+background* (51.4%) and *no depth* (46.8%). Finally, for the eight classes on which good performance is achieved, the 'Atm.scatter+perspective' feature set

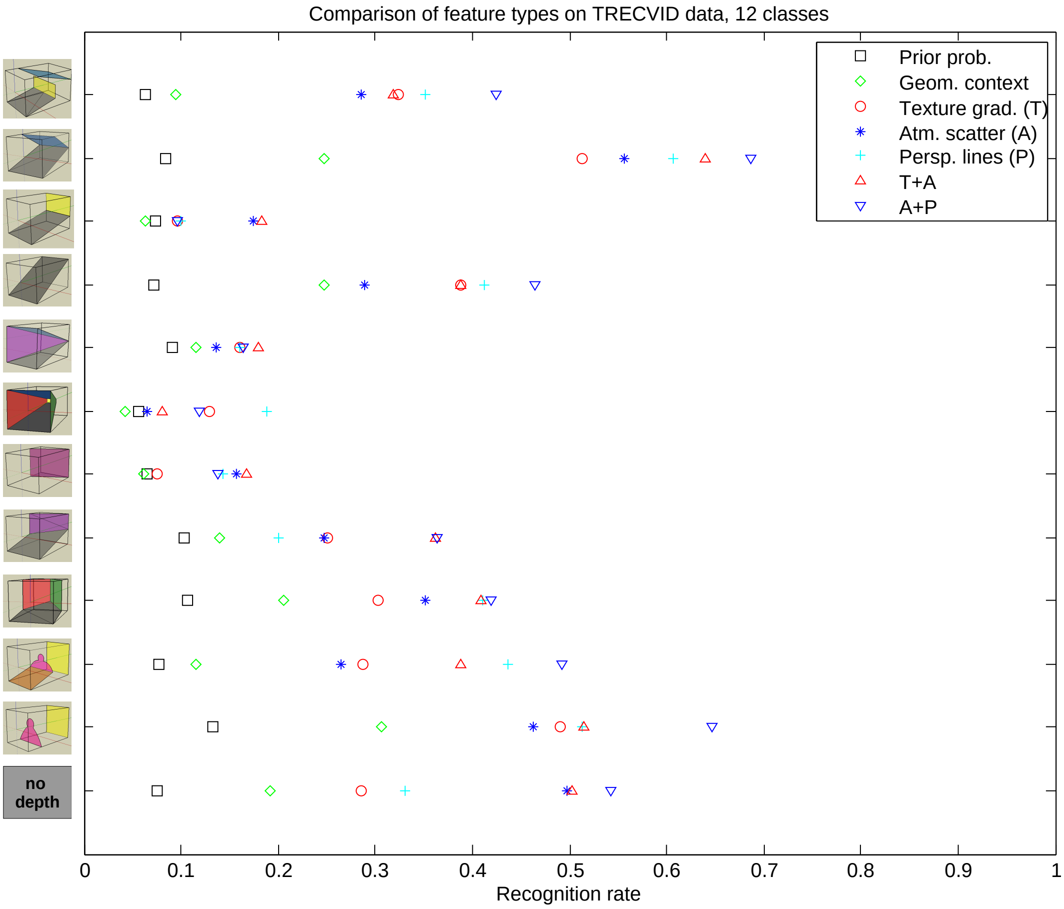
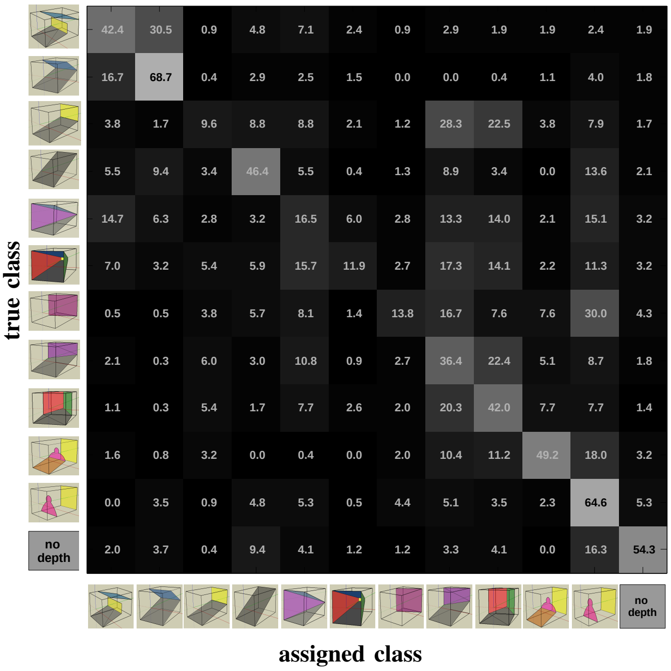


Fig. 9: Comparison plot of different feature sets in stage classification. Note that results with 'Geometric context' features have been obtained with AdaBoost classifier, whereas the rest is achieved with support vector machines.

performs best, reaching almost 70% recognition for stages *sky+ground* and *person+background*. In case of the poorly-recognized classes, texture gradients are more important than perspective lines, as these images are marked by vertical walls where other information may not add much value. The exception is the *corridor* stage, where perspective lines are distinctively better; this is expected, considering that the stage model contains tilted lines at four different image corners.

Additional insight into performance is acquired from the confusion matrix given in Figure 10, describing results with the best-performing 'Atm.scatter+perspective' feature set. Stages are mostly confused with geometrically similar types, or with those to which they have a direct link in the transition graph of Figure 3. In addition, the poorly recognized *ground+background* stage is mostly being mistaken for *ground+tilted background* and *corner* types; *tilted background* is confused with *person+background*; whereas *corridor* and *1 side-wall* stages are often exchanged with all those three types receiving many confusions. However, stages receiving lots of confusions are also the most likely ones in the dataset - *ground+tilted background* is represented with 10.3% of samples, *corner* with 10.7% and *person+background* 13.2%.

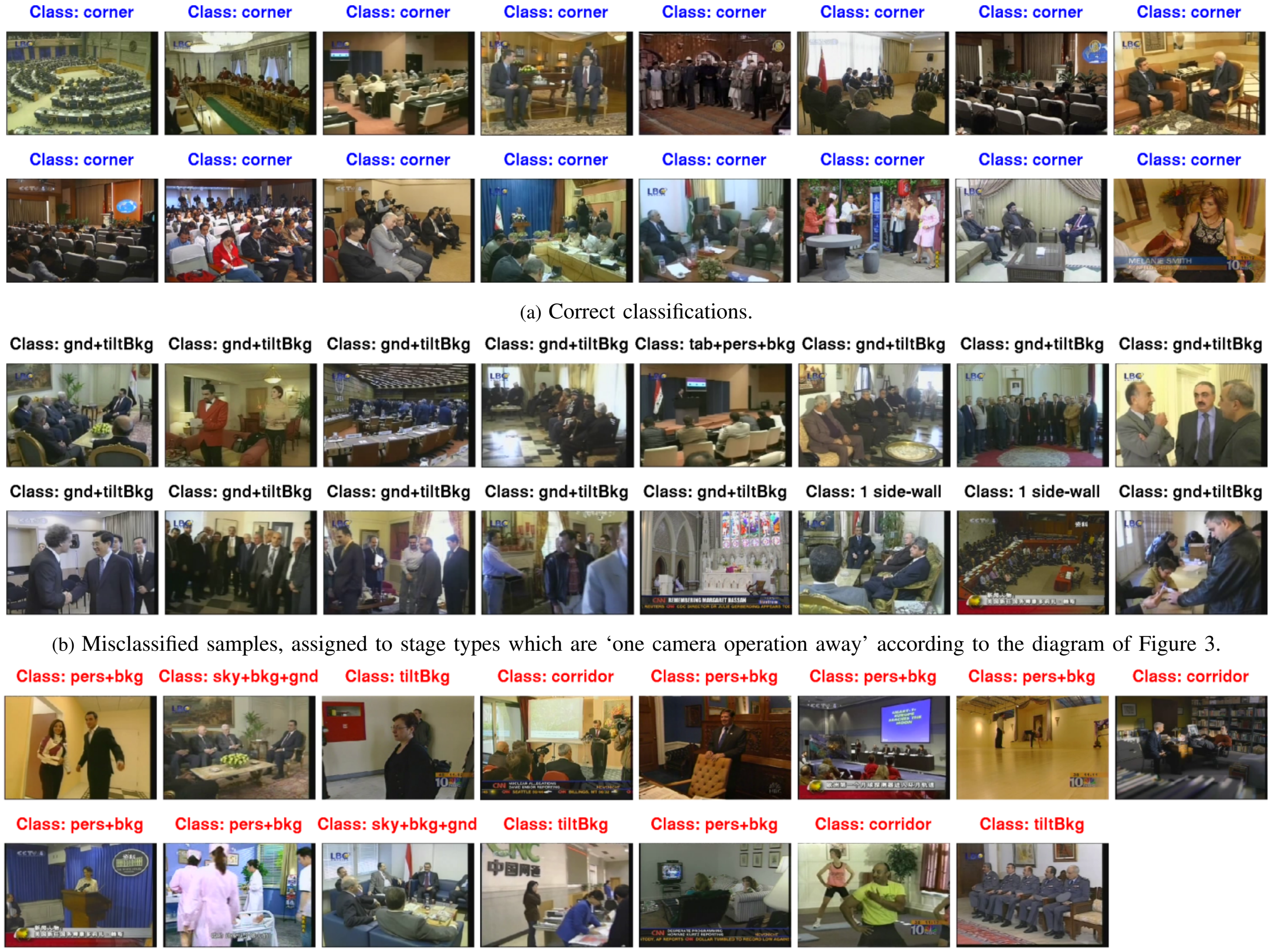
Classification results are demonstrated more clearly by an example of actual class assignment for the *corner* stage, with 'Atm.scatter+perspective' features. Figures 11a-11c show image samples and their classifications. Respectively, they



												no depth
	42.4	30.5	0.9	4.8	7.1	2.4	0.9	2.9	1.9	1.9	2.4	1.9
	16.7	68.7	0.4	2.9	2.5	1.5	0.0	0.0	0.4	1.1	4.0	1.8
	3.8	1.7	9.6	8.8	8.8	2.1	1.2	28.3	22.5	3.8	7.9	1.7
	5.5	9.4	3.4	46.4	5.5	0.4	1.3	8.9	3.4	0.0	13.6	2.1
	14.7	6.3	2.8	3.2	16.5	6.0	2.8	13.3	14.0	2.1	15.1	3.2
	7.0	3.2	5.4	5.9	15.7	11.9	2.7	17.3	14.1	2.2	11.3	3.2
	0.5	0.5	3.8	5.7	8.1	1.4	13.8	16.7	7.6	7.6	30.0	4.3
	2.1	0.3	6.0	3.0	10.8	0.9	2.7	36.4	22.4	5.1	8.7	1.8
	1.1	0.3	5.4	1.7	7.7	2.6	2.0	20.3	42.0	7.7	7.7	1.4
	1.6	0.8	3.2	0.0	0.4	0.0	2.0	10.4	11.2	49.2	18.0	3.2
	0.0	3.5	0.9	4.8	5.3	0.5	4.4	5.1	3.5	2.3	64.6	5.3
no depth	2.0	3.7	0.4	9.4	4.1	1.2	1.2	3.3	4.1	0.0	16.3	54.3

Fig. 10: Confusion matrix for the best performing 'Atm.scatter+perspective' feature set.

represent the correctly classified samples, misclassifications into stages that are 'one camera operation away' from the



(a) Correct classifications.

(b) Misclassified samples, assigned to stage types which are 'one camera operation away' according to the diagram of Figure 3.

(c) Misclassified samples, assigned to stage types unrelated to the ground truth model.

Fig. 11: Stage classification of *corner* images, with actual class assignment indicated above the image.

ground truth model, and misclassifications into unrelated classes. From the figures, the diversity of the dataset and the difficulty of the problem become apparent. Nevertheless, the classifier is able to recognize *corner* images in many cases when the clutter does not obstruct the view of stage walls. The majority of confusions are with classes that are geometrically related to the ground truth model, most of them being with the *ground+tilted background* class. There are, however, some misclassifications which do not result in similar geometry, but where occlusion, foreground objects, or image colors bias the decision.

4.5 Stage hierarchy

Given the continuum of camera parameters (Figure 3), and our stage models as the samples of that continuum, individual stages could be organized or grouped in several ways. This process would ideally result in a hierarchical ordering of models, which would be entirely based on geometry. In such a hierarchy, finer classification would impose more geometrical restrictions on the stage configuration.

Figure 12 shows one possibility of a stage hierarchy. By moving down in the figure, a more constrained geometry is obtained. It also presents a possibility of grouping individual stages into higher-level geometrical categories. That way, classification is also defined at the level of 6 stage groups (i.e. *super-stages*), shown by orange boxes in the figure.

We have evaluated this stage hierarchy by performing classification at the level of proposed stage groups. This is shown in Figure 13. Again there are certain classes, here named *straight/no background, person+background* and *no depth*, which are recognized well. All three can be detected in at least 49% of the cases, with performance on *straight/no background* reaching 72%. Only moderate performance is achieved for the other classes, which are usually the super-stages of poorly-recognized models in Figure 9. The 'Atm.scatter+perspective' feature set again performs best in most of the cases, with 'Weibull+atm.scatter' combination outperforming it occasionaly. Thus for classes *tilted background* and *no depth*, 'Texture+atm.scatter' is the best feature set, whereas for the *corridor* class perspective information by itself marginally improves the results.

A confusion matrix corresponding to the best performance, achieved with 'Atm.scatter+perspective' feature set, is shown in Figure 14. The confusions are similar to some degree to

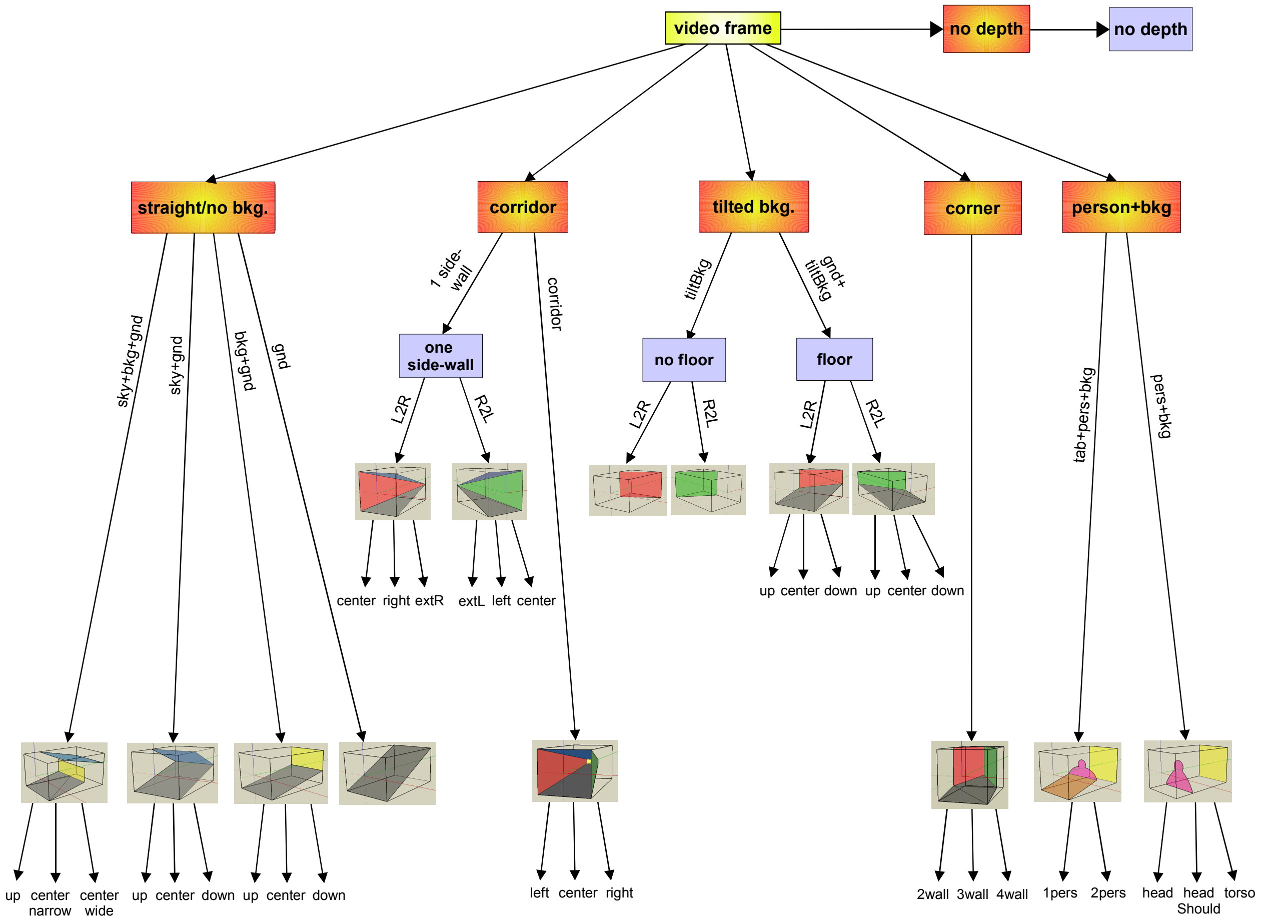


Fig. 12: A possible hierarchy of stage models, with 6 classes at the level of *super-stages* in addition to 15 original stages. Further geometrical restrictions could be placed on 3D models by the third level of division, which may be regarded as the rough stage parameter estimation.

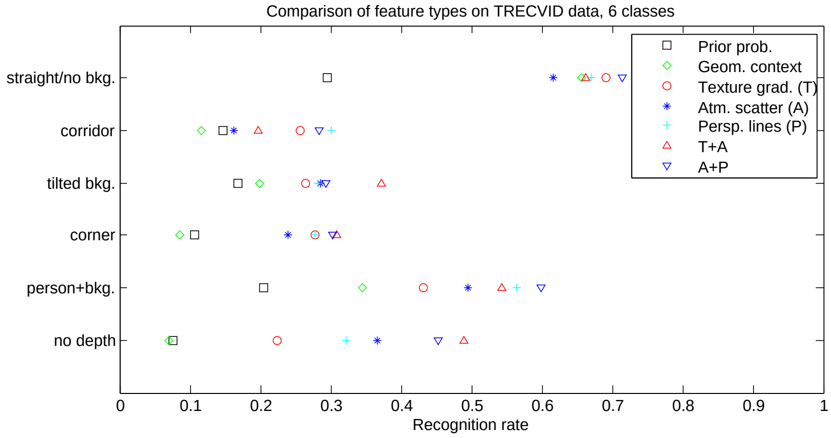


Fig. 13: Comparison plot showing the performance of different feature sets at the level of 6 stage groups.

those of individual stages. Namely, the *straight/no background* group is recognized in over 70% of the cases, although samples of other classes are often confused with it, accounting for high numbers in the first column of the matrix. Similarly, images are frequently incorrectly classified as belonging to *tilted background* and *person+background* groups. These stage groups, though, are the best represented ones in the dataset - *straight/no background* contains 29.5% of all samples, *tilted background* contains 16.8% and *person+background* 20.5%.

For completeness, Table 5 summarizes the performance of various feature sets, at both levels of the hierarchy defined by Figure 12. It gives average stage classification results for all the experiments described in this paper.

4.6 Independent dataset

In order to investigate how our approach generalizes to other image data, we have performed another set of experiments on the collection of our own 2007 television recordings. These recordings comprise samples from 14 broadcasting channels over 13 days, containing 110166 frames in Dutch (from

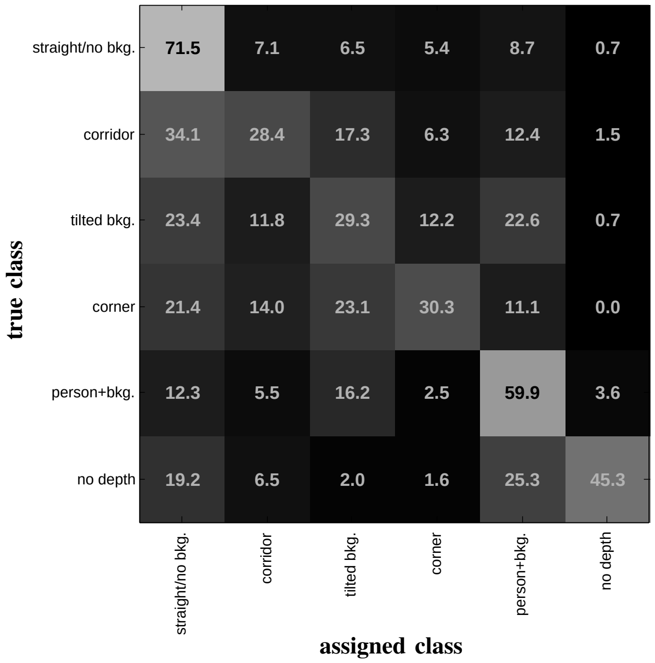


Fig. 14: Confusion matrix for the best classification results on 6 stage-groups - ‘Atm.scatter+perspective’ feature set.

Feature set	**Average performance**	
	12 stages	6 stage groups
Geom. context (960 f./AdaBoost)	**15.3**	**24.5**
Texture grad. (T) (64 f./SVM)	**27.5**	**35.8**
Atm. scatter (A) (80 f./SVM)	**29.1**	**36.1**
Persp. lines (P) (128 f./SVM)	**32.1**	**40.3**
T+A (144 f./SVM)	**34.4**	**42.9**
A+P (208 f./SVM)	**38.0**	**44.1**

TABLE 5: Average stage classification results with various feature sets at both stage levels.

channels Nederland 1, 2 and 3) and English (Al Jazeera, CNN, Discovery and Eurosport). A subset of 3551 such frames has been annotated and used for stage classification purposes[1].

Comparison plots for performance of different features on TV data are given in Figures 15 and 16. The analysis of Figure 15 results in similar conclusions as for the TRECVID dataset. Namely, in almost all of the cases, the same classes are recognized very well (e.g. the three last stages - *table+person+background, person+background* and *no depth* -

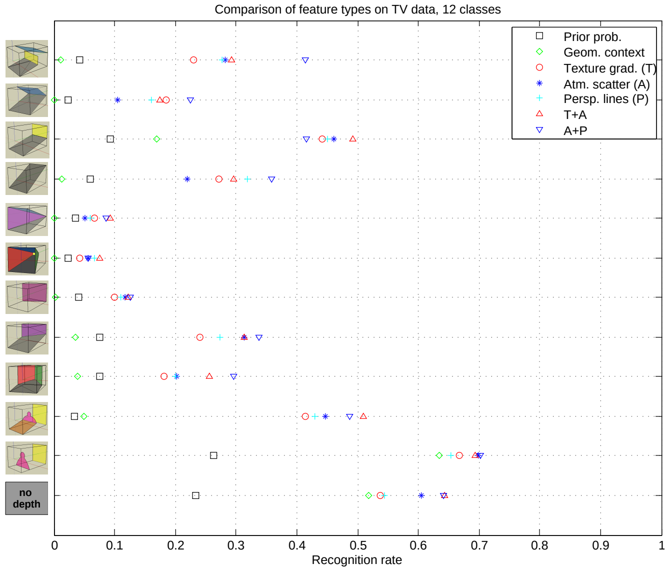


Fig. 15: Comparison plot showing the performance of different feature sets on TV data, at the level of 12 individual stages.

are detected with over 50% accuracy with the best performing feature set), and the same ones poorly (i.e. *1 side-wall* and *corridor* recognized in less than 10% of the cases). The exception are the stages *sky+ground* and *ground+background*, which seem to have exchanged places - in this dataset, the recognition rate of *sky+ground* drops to ∼25% (compared to 68.7% in TRECVID experiments), whereas the rate for *ground+background* is reaching almost 50% (relative to 18.3%). This is due to different distribution of confusions for the *ground+background* class. The ‘Atm.scatter+perspective’ feature set remains the best choice overall, except in case of *ground+background, corridor* and *table+person+background* stages, where ‘Texture+atm.scatter’ performs slightly better. In addition, except for ‘Geometric context’ features, most of the feature sets achieve significantly better results than baseline. Similar to the TRECVID dataset, the difference between prior probability and the best achieved result is the biggest for *table+person+background* (47.6%), *person+background* (43.9%), and *no depth* (40.9%) stages.

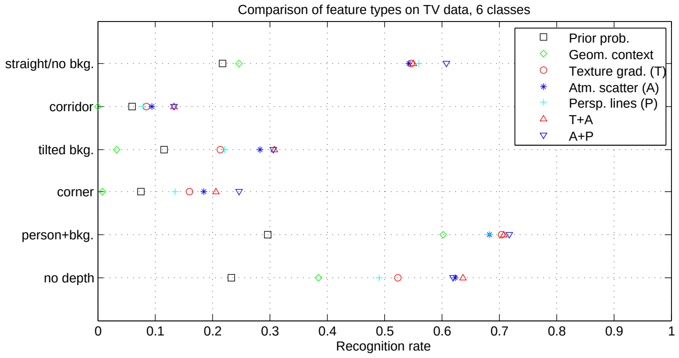


Fig. 16: Comparison plot showing the performance of different feature sets on TV data, at the level of 6 stage groups.

From Figure 16, showing classification of the TV recordings into 6 proposed stage groups, we again draw similar

[1] Due to copyright issues, we are unable to show example frames from this dataset in the paper.

conclusions as for the TRECVID data. Classes *straight/no background, person+background* and *no depth* are recognized with more than 60% accuracy with the best performing feature set, whereas moderate performance is achieved for other stage groups. The 'Atm.scatter+perspective' feature set remains the best choice in most of the cases, being slightly outperformed by 'Texture+atm.scatter' combination in case of the *no depth* group (containing only one member).

Overall, classification performance and conclusions derived with the independent dataset are very similar to those with TRECVID data, demonstrating the robustness of the proposed methods and features.

5 Discussion

The geometry of scenes changes very differently from the geometry of objects acting in them. Whereas the shapes of foreground objects may vary greatly, the scenes are more or less stable geometrical environments. This is due to many regularities present in images, arising both from inherent structure of the world and from constraints related to camera viewpoint. Since for many applications only a rough depth model suffices for the scene, the image structure can be accounted for by defining a limited number of 3D models - *stages* - for typical scene geometries.

Based on these observations, we present a method to infer weak scene geometry from a single image. We do so via stage classification. Each stage has a specific 3D profile and serves as the first approximation of global scene depth. In addition to providing a background depth model, stage information narrows down the possibilities with respect to objects' locations, scales and identities.

We investigate visual features relevant for depth estimation. Low-dimensional representations are consequently proposed, indicative of texture gradients, atmospheric scattering and perspective line information. For texture gradients and perspective lines, Weibull γ parameter shows measurable correlation with increase in overall depth, whereas β reaches the maximum value at the location of camera focus. We also conclude that saturation invariably decreases with depth, as is the case for color correction coefficient for the blue channel, regardless of the complex content of the scene. The color correction coefficient for the red channel is, on the contrary, negatively correlated with scene distances.

Using the proposed features, we obtain detailed quantitative stage classification results on the keyframes dataset of the 2006 TRECVID benchmark. Comparison of performance demonstrates that recognition rates achieved with the 'Geometric context' [10] feature set are always lower than with the proposed features. This is the case despite its high dimensionality, and despite the improvement achieved by using AdaBoost classifier instead of SVMs. Results with Weibull texture gradient features are somewhat poorer than those presented in [8]; this is due to the pre-processing steps described in Section 3.4, which inevitably remove some useful pictorial information from the image (e.g. if subtitles are detected, the whole bounding box is discarded). The combination of atmospheric scattering and perspective information, containing only 13 features in each of the 16 regions, emerges as the best choice for stage classification in general. However, the more compact set combining texture and atmosperic scatter, comprising 9 features per region, often approaches similar performance, making it a good choice when efficiency is a concern.

The results indicate that some simple scene configurations can be detected very robustly. This is true for classes which typically appear with small content variations (e.g. members of the *person+background* stage group), which represent open scenes (*straight/no background* group), or which are not likely to contain object clutter. In these cases, eight well-recognized stages can be detected with more than 35% accuracy, whereas 70% performance is approached for the *sky+ground* and *person+background* types. On some of the other stages, however, our detector is not performing as well. This is due to the lower number of learning samples, the amount of variation within the class, and the significant amount of occlusion and object clutter - as in *1 side-wall, corridor* and *tilted background* stages.

More detailed analysis of actual class assignments, as shown in Figure 11 for images of the *corner* category, demonstrates that most misclassifications arise from confusions with geometrically related stages, i.e. those which are 'one operation away' from the ground truth model in the stage transition graph of Figure 3. In fact, tolerating a deviation of one camera operation while computing classification performance leads to average recognition rate of 69% for the 12 stages, with highest accuracy of 90% in case of the *sky+background+ground* type, and lowest (31%) in case of the *corridor* stage.

Generic concept detection of video data remains a challenging problem. The performance in the TRECVID benchmark reaches some 40% recognition rate for concepts such as street, classroom, boat, and so on, after many rounds of performance upgrades over the last several years [40]. Here, using a much smaller feature set than in TRECVID systems, we achieve a similar average recognition rate of 38% among 12 different classes. This number is obtained regardless of real data with large diversity in topic, often polluted by accidental graphics. In addition, the results are confirmed by the evaluation on an independent dataset, demonstrating the robustness of the proposed approach.

References

[1] W. Richards, A. Jepson, and J. Feldman, "Priors, preferences and categorical percepts," in *Perception as Bayesian Inference*, D. Knill and W. Richards, Eds. Cambridge University Press, 1996, ch. 3, pp. 80–111.

[2] B. Horn and M. Brooks, *Shape from Shading*. MIT Press, 1989.

[3] L. Matthies, T. Kanade, and R. Szeliski, "Kalman filter-based algorithms for estimating depth from image sequences," *IJCV*, vol. 3(3), pp. 209–238, 1989.

[4] S. Barnard, "A stochastic approach to stereo vision," in *5th National Conference on Artificial Intelligence*, 1986.

[5] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *IJCV*, vol. 75(1), pp. 151–172, 2007.

[6] D. Hoiem, A. Stein, A. A. Efros, and M. Hebert, "Recovering occlusion boundaries from a single image," *ICCV*, 2007.

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

[8] V. Nedović, A. W. M. Smeulders, A. Redert, and J. M. Geusebroek, "Depth information by stage classification," *ICCV*, 2007.

[9] A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and TRECVid,” in *ACM MIR*, 2006.

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

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

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

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

[14] A. Saxena, S. Chung, and A. Ng, “3-D depth reconstruction from a single still image,” *IJCV*, vol. 76(1), pp. 53–69, 2008.

[15] E. Delage, H. Lee, and A. Ng, “A dynamic bayesian network model for autonomous 3D reconstruction from a single indoor image,” *CVPR*, 2006.

[16] Z. Yang and D. Purves, “A statistical explanation of visual space,” *Nature Neuroscience*, vol. 6(6), pp. 632–640, June 2003.

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

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

[19] H. G. Barrow and J. M. Tenenbaum, “Interpreting line drawings as three-dimensional surfaces,” *Artif. Intelligence*, vol. 17(1-3), pp. 75–116, 1981.

[20] A. P. Pentland, “Fractal-based description of natural scenes,” *IEEE PAMI*, vol. 6, pp. 661–674, 1984.

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

[22] 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.

[23] A. Vailaya, A. Jain, and H. Zhang, “On image classification: City images vs. landscapes,” *Pattern Recognition*, vol. 31(12), pp. 1921–1935, 1998.

[24] 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.

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

[26] 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*, 2005.

[27] A. Bosch, A. Zisserman, and X. Munoz, “Scene classification via pLSA,” *ECCV*, 2006.

[28] 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.

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

[30] J. Huang and D. Mumford, “Statistics of natural images and models,” *CVPR*, 1999.

[31] D. L. Ruderman and W. Bialek, “Statistics of natural images: Scaling in the woods,” *Physical Review Letters*, 1994.

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

[33] 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*, vol. 28, 2006.

[34] S. Narasimhan and S. Nayar, “Vision and the atmosphere,” *IJCV*, vol. 48(3), pp. 233–254, 2002.

[35] F. Cozman and E. Krotkov, “Depth from scattering,” *CVPR*, 1997.

[36] J. M. Geusebroek, A. W. M. Smeulders, and J. van de Weijer, “Fast anisotropic gauss filtering,” *IEEE Trans. on Image Processing*, vol. 12(8), pp. 938–943, 2003.

[37] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2000.

[38] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[39] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm,” in *Proc. of the 13th International Conference on Machine Learning*, vol. 148, 1996.

[40] C. G. M. Snoek et al., “The MediaMill TRECVID 2008 semantic video search engine,” in *Proceedings of the 6th TRECVID Workshop*, 2008.

PLACE PHOTO HERE

Vladimir Nedović Biography text here.

PLACE PHOTO HERE

Arnold W.M. Smeulders Biography text here.

PLACE PHOTO HERE

Jan-Mark Geusebroek Biography text here.

PLACE PHOTO HERE

André Redert Biography