Single image-based 3D scene estimation from semantic prior

Hyeong Jae Hwang and Sang Min Yoon✉

Reconstructing a three-dimensional (3D) structure from a single image sequence to provide relevant contextual information for better human visual perception is a fundamental problem in computer vision. A 3D scene estimation methodology from a segmented image sequence that is learned from semantic priors is proposed. In particular, semantic information including 3D geometric characteristics can very efficiently predict the 3D structure of the scene from a given semantic region. The approach, which utilises semantic priors to estimate a 3D scene, is very robust for direct 3D scene reconstruction from an ambiguous depth map. The efficiency and effectiveness of the proposed approach has been proven experimentally with a large database.

Introduction: Estimating a three-dimensional (3D) scene structure from a single image has received great attention from numerous researchers in the fields of robotics, surveillance, and general scene analysis. The main purpose of 2D-to-3D estimation methodologies is to provide relevant contextual information that is very similar to the human visual perception system. To effectively understand the structure of a 3D scene, previous major approaches utilised the techniques of stereovision, structure from motion, and other approaches, which began with the assumption that they already knew the geometric characteristics of the two images. Recently, single image sequence-based 3D scene analysis from unknown geometric priors was proposed by automatically predicting depth cues [1–3]. However, most single image-based 3D scene reconstruction approaches ignore the task of semantic understanding, and immediately estimate the depth or geometry from the unknown image [2, 3]. These approaches also started from the assumption that they already have camera intrinsic parameters and height of the camera. However, the reconstruction of a 3D scene from the directly predicted depth map of a single image is still crucial to resolving the depth ambiguities resulting from similar appearances, colour, and partial occlusion.

In this Letter, we propose a novel methodology of reconstructing a 3D scene structure, which estimates the geometric and semantic information from a given image by predicting the depth map from automatically extracted vanishing points and segmented labels taken from a large image repository. Fig. 1 illustrates the proposed 3D scene estimation process using segmented labels. The main contribution of this Letter is the vanishing point-based camera calibration and depth map estimation focused on estimating the geometric structure, and the segmented prior provides the semantic relationship between the labels. The incorporation of semantic features and depth with the geometric constraints allow us to achieve significant results.

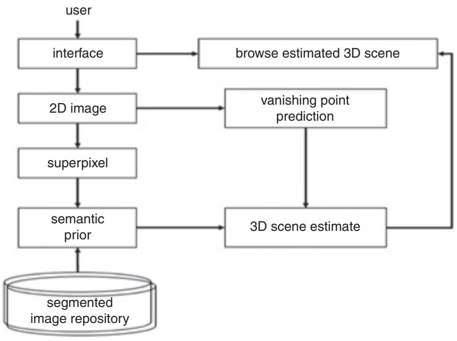
user
interface
browse estimated 3D scene
2D image
vanishing point prediction
superpixel
semantic prior
3D scene estimate
segmented image repository

Fig. 1 *Complete flowchart for estimating 3D scene from single image using semantic priors*

3D scene estimation by segmenting objects and predicting the vanishing point: The structure estimation of a 3D scene from an unknown image consists of two phases. The first step is to predict the semantic and geometric class of each pixel using superpixel-based labelling from a learned prior. Next, given this semantic and geometric context, we estimate the depth and structure of the 3D scene.

In the first phase, we formulate the region annotation of the given image I as superpixel-based labelling from the learned prior. Image segmentation and annotation can be computed efficiently by retrieving a relatively small set of training images that will serve as the source of the candidate superpixel level. To label the region which has characteristics similar to the top ranked images using an appearance feature-based matching procedure, we use a superpixel (s_i) which represents a restricted form of region segmentation, thus reducing image complexity through pixel grouping while avoiding under-segmentation. The goal of semantic and geometric segmentation is to assign an object label to each pixel, analogous to the general image parsing problem [4]. We present the set of labels $L \in \{l_i\}$, where $i \in U$ denotes a region index within a set of superpixels in I, and l_i denotes the label of a semantic and geometric region indexed by i. The similarity of shape, location, colour, texture, and appearance of the superpixel of the image I is measured from the high ranked images' superpixels which are already assigned by users according to the image's contextual and geometric information. For each superpixel and feature type, we find the nearest neighbour superpixels in the retrieved images using the log likelihood ratio, which is defined as

$$L(s_i, c) = \log \frac{P(s_i|c)}{P(s_i|\bar{c})} \tag{1}$$

where $\bar{c}$ is the set of all classes excluding c. We can obtain an image label by simply assigning each superpixel the class that maximises (1).

In the framework of semantic and geometric labelling, we assign the labels based on the most likely joint region label assignment $\hat{L}$ under a probability distribution $P(L|I) : \hat{L} \in \mathrm{argmax}_L P(L|\hat{X}, I)$ given by learned semantic priors, $\hat{X}$ such as sky, road, building, person, car etc. We segment and assign labels to superpixels, or regions produced by bottom-up segmentation to reduce the complexity of the problem and provide better spatial support for aggregation features that could belong to a single object, rather than fixed-size square patches centred on every pixel in the image. We model the probability distribution $P(L|\hat{X}, I)$ with a second order conditional random field

$$\begin{aligned} \ln P(L|X, I) = {} & \sum_{i\in U} \Phi(l_i|X, I) + \sum_{(i,j)\in V} \lambda_1 \Psi_1\big(l_i, l_j\big) \\ & + \sum_{(i,j)\in V} \lambda_2 \Psi_2\big(l_i, l_j|X, I\big) - \ln Z \end{aligned} \tag{2}$$

where V is a set of neighbouring pairs of image regions, λ_1 and λ_2 are model parameters, and Z is a partition function [5]. We model the unary potential function Φ using the probability of a label assignment, given the feature representation of the image region as $\Phi(l_i|X, I) = \ln P(\mathrm{l_i}|\phi(s_i, X))$. Fig. 2 shows the labelled image sequences from learned semantic priors of the retrieved image set. The labelled regions of the image can be separated according to semantic and geometric characteristics.

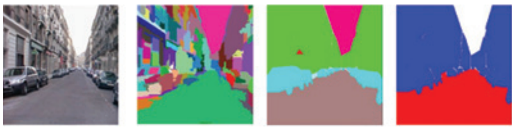

Fig. 2 *Semantic and geometric segmentation using superpixels from input image*

In the second phase, we then consider the depth from the predicted semantic and geometric segmentation. For the depth estimation of each pixel, we need to find a vanishing point of the image. Canny edge and Hough transform are used on a geometric segmented image to obtain a vanishing point [6, 7]. With the assumption that only one vanishing point exists on the image, we explore the points of intersection of lines produced by Hough transform and calculate their mean value, the vanishing point. Selecting critical lines by adjusting the number of votes of Hough transform enables us to find a stronger vanishing point. Then, when we find the points of intersection, we ignore near-vertical lines and do not explore the points of intersection of lines having similar directions.

The camera eye level equals the coordinate y of the predicted vanishing point. Therefore, we can learn the camera angle from the position of the vanishing point on the image. From estimated vanishing point, we compute the depth (d) from given image as shown in Fig 3. The depth value of the point in the given image is calculated from camera angle according to the vanishing point. The depth value can be expressed as follows:

$$d_y = \frac{bl_i h_c}{h_i} \tag{3}$$

where h_c is the camera height on the ground, bl_i is the camera inner base line which is computed by $bl_i = \sqrt{f^2 + (H/2 - p_y)^2 - h_i{}^2}$, h_i is the camera inner height $h_i = (vp_y - p_y)f/\sqrt{f^2 + (vp_y - H/2)^2}$, and H is the image height. We assume that the focal length f and the camera height h_i equal H. As shown in (3), we can successfully estimate depth value using vanishing point and height of the camera position (h_c).

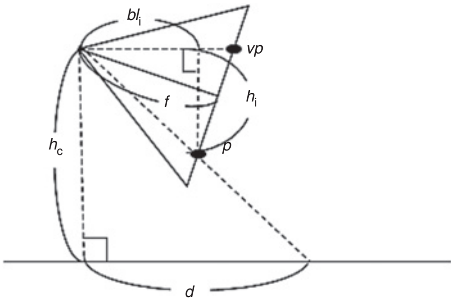

Fig. 3 *Illustration to compute depth from camera centre*

From estimated vanishing point, we construct the cuboid space for each segmented object in a given image. Then, we need to ensure that the object settles down into its cuboid. Without considering objects floating in the air, we can see at most three sides (the front, the upper, and the side). The front and the side of the objects follow the vanishing lines to effectively visualise in the estimated 3D scene structure.

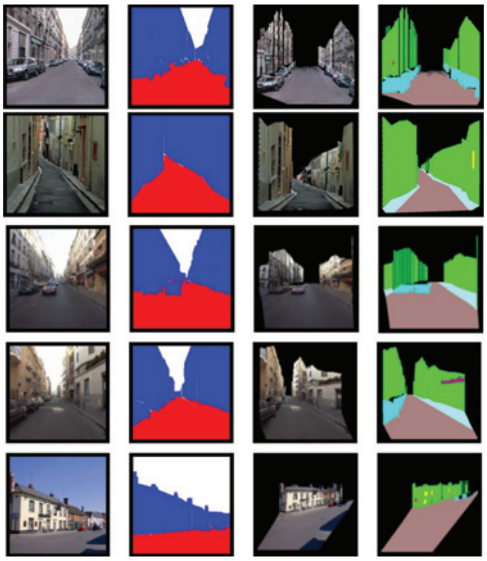

Fig. 4 *Predicted 3D structure of scene from single image by using segmented semantic and geometric prediction*

Experimental results: We evaluated the performance of our proposed 3D scene estimation on a single image from Tighe and Lazebnik dataset [4]. The dataset consists of 5300 images with corresponding semantic and geometric segmentation into 5100 training images and 200 test images. To retrieve the relatively similar images from a large database, we used several global feature extraction methodologies, such as pyramid, GIST, and colour histogram. The top ranked 100 image candidates were used to find the semantic and geometric characteristics from Naïve Bayes classifier. Fig. 4 represents the estimated 3D scene from the input image by predicting its semantic segmentation and depth. Even though the vanishing point is not defined from the input image, we successfully estimated the depth map by using vanishing point and the structure of the geometric and semantic priors. Fig. 4 also represents the structure of the 3D scene captured from various areas. As shown in Fig. 4, the 3D structure of the scene from a single image can be successfully recovered even though the vanishing point is not estimated within a given image using only an estimated semantic and geometric labelling procedure. The semantic and geometric segmentation can be visualised in the estimated 3D structure of the scene with different colours.

Conclusion: We presented an algorithm for estimating the 3D scene structure from a single image by learning its semantic prior and predicting vanishing point without any prior information about the camera's intrinsic parameters. Compared with previous approaches, our methodology provided a more accurate 3D scene model from segmented semantic and geometric labelling from a retrieved image set from a large repository. The semantic information of the given image was estimated from the superpixel and retrieved images' semantic prior in the large repository. The 3D structure of the scene could be estimated from single image without any prior information. We aim to extend our approach to estimate the entire 3D scene and its predicted semantic prior and focal length from simultaneously updated image sequences.

Acknowledgments: This research was supported by the CRC program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science (NRF-2015R1A5A7037615). It was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1002890)

Submitted: *27 April 2015* E-first: *8 October 2015*
doi: 10.1049/el.2015.1458
One or more of the Figures in this Letter are available in colour online.

Hyeong Jae Hwang and Sang Min Yoon (*HCI LAB., School of Computer Science, Kookmin University, 77 Jeongnueng-ro, Sungbuk-gu, Seoul 136702, Republic of Korea*)

✉ E-mail: sangmin.yoon@gmail.com

References

1 Hoiem, D., Efros, A.A., and Hebert, M.: 'Recovering surface layout from an image', *Int. J. Comput. Vis.*, 2007
2 Saxena, A., Sun, M., and Ng, A.Y.: 'Make3D: learning 3D scene structure from a single still image', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008
3 Lin, D., Gould, S., and Koller, D.: 'Single image depth estimation from predicted semantic labels'. Proc. of CVPR, 2010, pp. 329–332
4 Tighe, J., and Lazebnik, S.: 'SuperParsing: scalable nonparametric image parsing with superpixels'. Proc. of ECCV, 2010
5 Malisiewicz, T., and Efros, A.A.: 'Recognition by association via learning per-exemplar distance'. Proc. of CVPR, 2008
6 Moghadam, P., and Dong, J.F.: 'Road direction detection based on vanishing point tracking'. Proc. of IROS, 2012
7 Rogers, S.D., Kadar, E.E., and Costall, A.: 'Gaze patterns in the visual control of straight-road driving and braking as a function of speed and expertise', *Ecological Psychol.*, 2005, **17**, pp. 19–38