21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Stage-Based 3D Scene Reconstruction from Single Image

Yixian Liu, Pengwei Hao, Ebroul Izquierdo
Queen Mary University of London
{yixian.liu;phao;ebroul.izquierdo}@eecs.qmul.ac.uk

Abstract

Holistic scene understanding is a major goal in recent research of computer vision. To deal with this task, reasoning the 3D relationship of components in a scene is identified as one of the key problems. We study this problem in terms of structural reconstruction of 3D scene from single view image. Our first step concentrates on geometrical layout analysis of scene using low-level features. We allocate images into seven recurring and stable geometry classes. This classification labels the image with rough knowledge of its scene geometry. Then, based on this geometry label, we propose an adaptive autonomous scene reconstruction algorithm which adopts specific approaches particularly for different scene types. We show, experimentally, given the right geometry label, low-quality uncalibrated monocular images from the benchmark dataset can be structurally reconstructed in 3D space in a time/effort efficient way. This robust approach does not require high quality or high complexity input image. We demonstrate the effectiveness of this approach in this paper.

1. Introduction

Recent years, much research has been made to let computer understand digital image as humans. In this paper, we study the reconstruction of basic scene structures from monocular images. As regards to the images shown in Figure 1, human beings can perceive depth and semantic information (background, ground, sky, streets, buildings, people, etc.) very easily by simply viewing the images. Human visual system functions in a non-mechanical way so that people can actually understand the scene. In the field of computer vision, we are trying to computerise this process to empower computer with similar ability. The ultimate goal is to make machine be able to interpret and convey all kinds of information as human does.

Normally, we interpret an image from two aspects – capturing objects and recognising the environmental frame in which objects can move around. By conducting these activities simultaneously, we could understand an image. Some classic depth estimation

Figure 1 Examples to illustrate stage types

methods focus on object recovery and normally demand high quality images or only estimate local information [3]. And some researchers try to interpret a scene from these two aspects in conjunction [1, 7, and 8]. However, objects are enormously changeable. They are too complicated to be estimated accurately. So object detection could hardly be helpful to scene recognition. In terms of scenes, if images are categorized into semantic or appearance scenarios, similar overlapping and deficiency of vocabulary problem could happen again [9]. On the contrary, the types of scene geometrical structures are relatively limited [2].

Recently, successful research has been done to recognise a scene by its spatial structure. Layout of the surfaces that constructs a scene is considered to be independent of semantic content information. Surface layout recovery by statistical learning [10] and scene geometry type classification by image statistics are two good examples [2].

Inspired by this research path, we analyse image by just focusing on the scene itself and omitting detailed objects (later, objects can be put back into the reconstructed 3D space as they were). We propose an algorithm that is adaptive to seven different scene geometries for their reconstruction (sufficient to cover a large proportion of image appearing on Internet or broadcasting[2]). Identifying the rough geometry provides us a basic but valuable depth profile of the scene. First, we classify general images into seven typical geometry models. Then, based on the structure

profile of each model, specific mechanisms are designed adaptively to rebuild different kinds of scenes by flat planes in a 3D space. For instance, colours are strong cues to reconstruct open view scenes while line segments are important to reconstruct box-structured scenes. Consequently, this approach can be applied to a really wide scope of images. No specific scenarios are restricted. Moreover, the input image capturing device and image quality are far less demanding than many other approaches. The aim is to efficiently create a rough immersive ambience for the low-end user by very limited, low-level resources that they can afford to provide. Nowadays enormous amount of end-user generated content is uploaded to personal/social network spaces in every minute. Our work could be quite suitable to deal with these uncalibrated, low quality target images and produce reasonable results.

The rest of this paper is organised as follows: scene geometry classification scheme is described in section 2 and section 3 gives specific approaches to do geometry based scene reconstruction. Then, some initial results are shown in section 4.

2. Image geometrical classification

Stage models are rough geometries used to classify scenes into limited typical types. One thing should be noticed is the concept of a scene. Most efforts of this work dedicatedly deal with environmental scenes. Images showing a view from certain distance (usually from 5 meters) to the camera or scattering views are valid contents. Generally, objects should be able to move around within the scene. No-depth shots for fine object or texture are less considered and recognised as close-up stage since they are unable to show a spatial structure. Interviewer close-up shots appeared in broadcasting dataset are categorised as close-up stage as well. Thus, the categories defined in our system are very comprehensive. We could include more possible inputs. (We are using broadcasting benchmark data for testing and think it is close enough to social network user generated data than many other datasets).

Nedovic et al. [2] stated the visual world gives 15 typical scene geometries. Due to its complexity, we revise the typical scene geometries and categorize them into only 7 functional stages in this research. Figure 1 shows six of them. The missing one is the above mentioned close-up stage. We give approximate planar models as their specific depth profiles and describe it by some semantic vocabularies. This gives a sense of what range of information we can get by simply performing stage classification.

Generally, each image shows its uniqueness more from local details. Though images are presented in different ways, their global statistical characteristics may be very similar. For example, two pictures are likely to appear different thoroughly while their colour histograms are exactly the same. Since distinguishing information is generally obtainable better from local statistical analysis, pre-process is involved here. We divide an image into 4×4 grid regions. The division is done uniformly, resulting (W/4) × (H/4) pixels in each region. (W and H indicate the amount of pixels horizontally and vertically.) Then, all the features are considered and compared region-wide. In this phase of the work, we treat and value every region equally importantly. Same features are extracted from each region and the results extracted from all 16 regions are concatenated into one vector for training and classification.

In the state-of-the-art works, five feature sets have been investigated and no dominating ones have been found. We learned from their results and made one multi-dimensional feature vector to do the work. Different features are normalized before they are combined together. Experiments are performed to evaluate the effectiveness and our result is compared with Nedovic's result in section 4.

3. Adaptive 3D reconstruction

The output of stage classification gives us knowledge about the rough geometry of a given image. Utilizing such information, we design the reconstruction algorithm according to the available depth profiles. As we have mentioned above, general scenes are divided into seven categories including five open view (no vertical boundaries) stages, one box-structured stage and one close-up/no-depth stage. In this paper, we propose open view stages' reconstruction approaches and we use D. C. Lee's corner detection technique [5] for box-structured stage reconstruction to complete the whole algorithm.

For open view scenes, boundaries of conjunct surfaces are keys for reconstruction. Scene geometry tells us how many boundaries need to be found and the postures of them (direction and correlation with

Table 1 Formula of stage types

Type	Model
1.gnd	$s_{1i} = c$
2. sky+bkg	$s_{2i} = f(b_{1i})$
3. bkg+gnd	$s_{3i} = f(b_{2i})$
4. sky+gnd	$s_{4i} = f(h_i)$
5. sky+bkg+gnd	$s_{5i} = f(b_{1i}, b_{2i})$

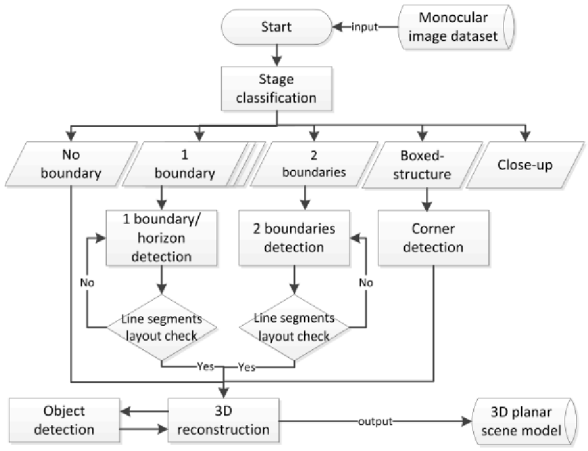

Figure 2 Flow chart of the whole process of the adaptive reconstruction algorithm

surfaces). For example, to build "bkg+gnd" stage, we need two planes and one junction line, and, to build "sky+bkg+gnd" stage, we need three planes and two junctions. Actually, to state the problem in an easier way, we can recover the stage structure as long as we know the exact position of the junctions. In our algorithm, we call the junction of sky and background planes the 1st boundary $b_1(\theta,y)$, the junction of background and ground planes the 2nd boundary $b_2(\theta,y)$, and the junction of sky and ground planes horizon $h(\theta,y)$, where θ indicates the tilted angle and y indicates its centre vertical position. Then, we formulate the reconstruction problem for these five open view stages in Table 1.

Now, our problem has been simplified into the task of finding b_1, b_2 and h. Horizon detection is a conventional topic in computer vision. We follow the principle described by Ettinger *et al.* in [11] with minor changes. Unlike the aircraft navigating devices, we assume that the tilted angle of the horizon $\theta_h \in [-10°, 10°]$ and the centre vertical position $y_h \in (1, H)$. The optimization criterion is using colour as measurement of appearance and described as

$$J = \frac{1}{|\Sigma_s|+|\Sigma_g|+|\lambda_1^s+\lambda_2^s+\lambda_3^s|+|\lambda_1^g+\lambda_2^g+\lambda_3^g|} \quad (1)$$

where, Σ_s and Σ_g are the covariance matrices of pixel distributions for sky plane $\Sigma_s = \frac{1}{(n_s-1)}\sum_{i=1}^{n_s}(x_i^s-\mu_s)(x_i^s-\mu_s)^T$ and ground plane $\Sigma_g = \frac{1}{(n_g-1)}\sum_{i=1}^{n_g}(x_i^g-\mu_g)(x_i^g-\mu_g)^T$, λ denotes the eigenvalues of Σ_s and Σ_g, and x_i, μ denote RGB values of the i-th pixel, global average respectively. Firstly, we select the optimal vertical position $(0, y_h^*)$ which minimizes the optimization criterion J. Then, to accurate the result, we find the pair (θ_h^*, y_h^*) to minimize J by choosing θ from the discrete set $\theta_h^i = (-10+2i)°, 0 \le i \le 10$.

We detect boundaries in a way similar to horizon detection approach. In addition, we take the result of line segment detection, getting by Hough transform, into consideration as well. For one boundary stage types, we use the above optimization criterion straightforwardly to get the boundary position. For the stage where two boundaries appear, we calculate J_1 and J_2 respectively and choose a winning pair (b_1^*, b_2^*) to minimize the sum of J_1 and J_2, $min(J_1+J_2)$. We check the result by line segments layout. Resulted boundary line should lie somewhere in the neighborhood of local cluster maximum. If the results match, we continue the reconstruction process and if not, we delete the position of this boundary from the pool and repeat from the top.

To reconstruct the scene in a 3D space, we simply put the pixels into the plane where they belong like folding papers. Simple object detection can be applied to each plane and pop-up the objects according to the nature of the plane (background – parallel pop-up/ground – vertical pop-up). The whole process of our work is described more clearly with as Figure 2.

4. Experimental results

We use key-frames from 2008 TREC Video Retrieval Evaluation (TRECVID) benchmark which consists of a wide scope of content from real broadcasting and surveillance videos. Approximately 300 images make up the experiment dataset and each of them belongs to one of the seven classes.

We adopt 1-vs-1 SVM hierarchically. The first-level classification includes three branches – open view (gnd, sky+bkg, bkg+gnd, sky+gnd, sky+bkg+gnd), box-structured view and close-up view. On the second level, open view images are further divided into stages 1-5, as numbered in Table 2. 1-vs-1 SVM is performed at each level. The classification results are shown in Table2. This result is comparable with Nedović's work.

Table 2 Accuracy ratio of the classification

Type	Classif. results	[2]'s results	Reconst. results
1.gnd	0.39	0.44	1.0
2.sky+bkg	0.46	n/a	0.62
3.bkg+gnd	0.20	0.17	0.53
4.sky+gnd	0.43	0.61	0.66
5.sky+bkg+gnd	0.29	0.17	0.41
6.box	0.35	0.25	0.7
7.close-up	0.46	0.38	n/a
Avg.	0.37	0.34	0.66

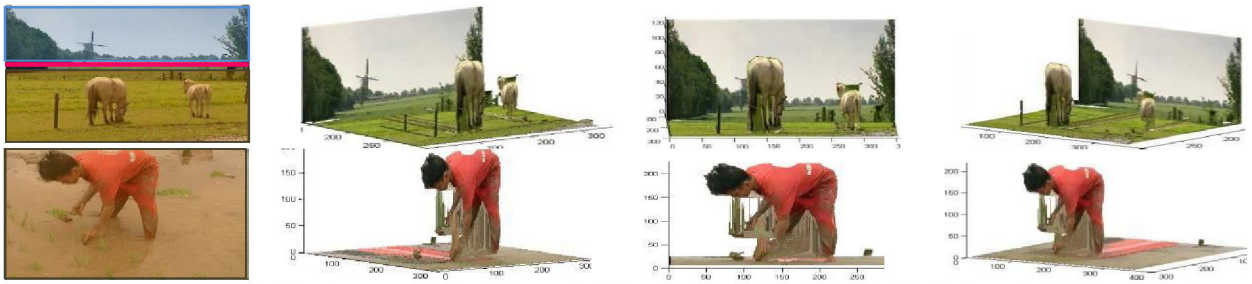

Figure 3 Examples of 3D planar scenes (from left to right: original, tilted left, front, tilted right views)

They are using 1-vs-1 SVM directly to the whole dataset. We have simplified the complexity of classes and the classification strategy by its hierarchy structure.

Then we reconstruct the scene according to their stage type label. Some results are shown in Figure 3. The top row shows the reconstruction of a "sky+ground" scene which is one of the 1-boundary stages. So we do single boundary detection. The pink line indicates the horizon that we found based on colour information. We fold the image in a concave way according to this horizon line, getting one sky plane and one ground plane. Then we do object detection on both planes. Those objects detected on the ground plane, we vertically pop them up (two horses, one bar and some small plants). The second row shows the reconstruction of a ground scene which contains no boundary. So we jump the boundary detection phase and do object detection right away. We fold the objects vertically, compensate the vacancy parts in the ground plane with the objects' neighbour ground pixels and make the ground surface lying horizontally.

The last column of Table 2 lists the accuracy rate of reconstruction. On average, 66% of the images returned acceptable 3D models. The inaccuracy results of this approach are mainly caused by the mislabeling of both stage type and boundary positions. Further improvements need to be investigated.

5. Conclusion and future work

Unlike many other 3D rebuild research works, the ultimate goal of our work is to recreate an immersive environment using content generated by low-end users. A novel approach to achieve this is proposed. At this phase, experimental results have shown this aim is achievable. The proposed adaptive autonomous scene reconstruction algorithm is efficient, well-rounded and robust for a large scale of image content. In the future, we will focus on improving the accuracy rate and trying to use more complicated and precise scene geometry models, to complete the algorithm more perfectly.

6. Reference

[1] D. Hoiem, A.A. Efros, and M. Hebert, "Geometric Context from a Single Image," Proc. IEEE Int'l Conf. Computer Vision, 2005.

[2] V. Nedović, A. W. M. Smeulders, A. Redert and J. Geusebroek, "Stages as models of scene geometry," IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), Vol. 32 (9), pp. 1673-1687, 2010.

[3] R. Zhang, P.-S. Tsai, J. Cryer, "Shape-from-shading: a survey," IEEE Transactions on Pattern Matching and Machine Intelligence, 21 (8) (1999), pp. 690–706.

[4] David C. Lee, Abhinav Gupta, Martial Hebert, and Takeo Kanade. "Estimating Spatial Layout of Rooms using Volumetric Reasoning about Objects and Surfaces." Advances in Neural Information Processing Systems 24 (NIPS) 2010.

[5] David C. Lee, M. Hebert, and T. Kanade, "Geometric Reasoning for Single Image Structure Recovery," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), June, 2009.

[6] E. Delage, H. Lee, and A. Y. Ng. A dynamic Bayesian network model for autonomous 3d reconstruction from a single indoor image. CVPR, IEEE Computer Society Conference on, 2:2418-2428, 2006.

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

[8] 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.

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

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

[11] S. Ettinger, M. Nechyba, P. Ifju and M. Waszak. "Vision-Guided Flight Stability and Control for Micro Air Vehicles", Conference on Intelligent Robots and Systems, Switzerland, Oct 2002.