Joint Spatio-temporal Depth Features Fusion Framework for 3D Structure Estimation in Urban Environment

Mohamad Motasem Nawaf and Alain Trémeau

Laboratoire Hubert Curien UMR CNRS 5516
Université Jean Monnet, Saint-Etienne, France
firstname.[midname.]lastname@univ-st-etienne.fr

Abstract. We present a novel approach to improve 3D structure estimation from an image stream in urban scenes. We consider a particular setup where the camera is installed on a moving vehicle. Applying traditional structure from motion (SfM) technique in this case generates poor estimation of the 3d structure due to several reasons such as texture-less images, small baseline variations and dominant forward camera motion. Our idea is to introduce the monocular depth cues that exist in a single image, and add time constraints on the estimated 3D structure. We assume that our scene is made up of small planar patches which are obtained using over-segmentation method, and our goal is to estimate the 3D positioning for each of these planes. We propose a fusion framework that employs Markov Random Field (MRF) model to integrate both spatial and temporal depth information. An advantage of our model is that it performs well even in the absence of some depth information. Spatial depth information is obtained through a global and local feature extraction method inspired by Saxena et al. [1]. Temporal depth information is obtained via sparse optical flow based structure from motion approach. That allows decreasing the estimation ambiguity by forcing some constraints on camera motion. Finally, we apply a fusion scheme to create unique 3D structure estimation.

1 Introduction

Estimating the 3D structure of a scene from 2D image stream is one of the most popular problems within computer vision. It is referred to as structure from motion (SfM) or 3D reconstruction from video sequence [2]. SfM has been applied in several applications [2] such as robot navigation, obstacle avoidance, entertainments, driver assistance, reverse engineering and modelling, etc.

In our work, we focus on the problem of estimating the 3D structure from a video taken by a camera installed on a moving vehicle in urban environments. This setup leads possibly to create 3D maps of our world. However, the dominant forward motion of the camera from one side, and the texture-less scenes that are present generally in urban environment produce an erroneous depth recovery. The forward camera motion could result degenerated configurations

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 526–535, 2012.

for a naturally ill-posed problem, or mathematically, a large number of local minima during the minimization of the reprojection error [3], that results in inaccurate camera relative motion estimation. Moreover, the limited lifetime of tracked feature points prevents using general optimization methods such as in traditional SfM. Additionally, forward motion restricts features matching due to non-homogeneous scale changes of image objects, especially those aligned parallel to camera movement.

In the proposed method, we suggest to benefit from the monocular cues (e.g. spatial depth information) to improve 3D depth estimation. We believe that such spatial depth information are complementary to temporal information. For instance, given a blue patch located at the top of an image, an SfM technique will probably fail to compute the depth due to the difficult matching problem, while the monocular depth estimation method (supervised learning) will assign it a large depth value as it will be considered as a sky with high probability.

Similar to other works [1][4][5], we consider that our world is made up of small planar patches, and the relationship between each two patches is either connected, planar or occluded. Based upon these considerations, the goal is to estimate the plane parameters where each patch lies. These patches are obtained from the image using over-segmentation method [6] or what is called superpixels segmentation. In order to fuse both temporal and monocular depth information, and also to handle the interactive relationship between superpixels, we proposed to use an MRF model similar to the one used in [1]. However, we extend the model by adding new terms to include temporal depth information computed using a modified SfM technique. Moreover we benefit from the limited Degrees of Freedom (DoF) of camera motion (which is such of the vehicle) to improve relative motion estimation, and in return, the depth estimation.

Spatial depth information is obtained using an improved version of the method proposed in [1], which estimates the depth from a single image. The method employs an MRF model that is composed of two terms; one integrates a broad set of local and global features, while the other handles the neighbouring relationship between superpixels based on occlusion boundaries. In our method, we compute occlusion boundaries from motion [7] to obtain more reliable results than using a single image as in the aforementioned method. Therefore, it is expected to have better reconstruction, even before integrating the temporal depth information.

To perform SfM, which represents temporal depth information, we use optical flow based technique that allows forcing some constraints on camera motion (which has limited DoF). Moreover, it is proved to have better depth estimation for small baseline distances and forward camera motion [6]. Here, we compute a sparse optical flow using an improved method of Lucase-Kanade with multiresolution and subpixel accuracy, results are refined thanks to camera motion estimation. Based on the famous optical flow equation [6], we obtain the depth for set of points in the image. Hence we can add some constraints on the position of scene patches to whom these points belong.

Outline of the Paper: In section 2 we give an overview of various methods for depth estimation using; video sequences, single image, and then combined spatio-temporal methods. In section 3 we introduce the MRF model that integrates SfM with the monocular depth estimation, and we explain its potential functions, parameters learning and inference. In section 4 we conclude our work and we discuss the advantages of the proposed method.

2 Related Works

In computer vision, structure from motion (SfM) has taken a great attention by researchers, it is considered as one of the well-studied problems. However, most of the efforts are focused on a certain number of aspects. For instance, improving feature points matching [8], formalize better constraints to improve relative camera pose estimation [9][10], robust methods for outliers rejection [11], linear/non-linear reprojection error optimization and bundle adjustment [10], formalize a set of constraints on more than two frames [9]. Most of these contributions do only consider temporal information that results from image stream variation with respect to time, without trying to analyse the monocular depth cues that are present in every single image.

From another side, several monocular cues that exist in a single image have been exploited by researchers, that includes; vanishing points and horizon line [9], shades, shadows, haze, patterns and structure [12]. Unfortunately, most of these cues are not present in all kinds of images, and they require specific settings. In contrast, we are looking to provide a general spatial depth estimation approach to be integrated with temporal depth estimation as mentioned earlier. Hence, we target a new generation (since last decade) of methods that perform 2D to 3D conversion using a single image. Generally these methods have no constrains and are based on the use of exhaustive feature extraction and probabilistic models to learn depth. An early approach attempts to estimate general depth of an image is proposed in [13] which employs Fourier spectrum to compute a global spectral signature of a scene to estimate the average depth of the image scene. Later on, an innovative attempt to perform 3D reconstruction from one image is proposed in [14]. Where first the image is over-segmented into superpixels, then each superpixel is classified as ground, sky or vertical. It employs a wide set of colour, texture, location, shape and edge features for training. Finally, the vertical region is “cut and folded” in order to create a rough 3D model. Although this method has been improved later by considering some geometric subclasses (centre, left, right, etc.) [15], the “ground-vertical” world assumption does not apply for wide range of images. More general method is proposed in [5] which estimates the depth from a single image based on some predicted semantic labels (sky, tree, road, etc.) using multi-class pixel-wise image labelling model. Then, the computed labels guide the 3D estimation by establishing a possible order and positioning of image objects. Another general approach has been proposed in [1] which does not have initial assumption about scene’s structure. It proceed by over-segmenting the image similar to [14]. The absolute depth of each image patch is estimated based on learning an MRF model, where a variety of features that capture local and contextual information is employed. We see later how a part of our work is inspired by this method.

In the context of combining both spatial and temporal depth information (as we aim), a method that combines SfM with a simultaneous segmentation and object recognition is proposed in [16], it targets road scene understanding. The task is achieved through a conditional random field model (CRF) which consists of pixel-wise potential functions that incorporate motion and appearance features. The author claims that it overcomes the effect of small baseline variations. In our method, we perform direct depth estimation rather than object recognition. However, similar to [16], our method is also supervised learning oriented, we benefit from computed features to capture contextual information and learn depth. Another approach with the same context is a semantic structure from motion approach [17] which is based on a probabilistic model. The proposed model incorporates object recognition with 3D pose and location estimation tasks. Also it involves potential functions that represent the interaction between objects, points and regions. In comparison with our approach, we use small planar patches [4] to model the world rather than the pixel-wise approach used in [16] as we think they better describe the world around us. Our idea is also supported by the experimental results in [18].

3 Spatio-temporal Depth Fusion Framework

In this section, we first introduce some notations. Then we explain how we compute spatial and temporal depth features. After that, we discuss estimating occlusion boundaries that play an important role in our model. Next, we introduce our proposed framework as an MRF model that incorporates several terms related to spatial and temporal depth features. Finally we show how we estimate the parameters from a given dataset and perform the inference for a new input.

3.1 Image Representation

As mentioned earlier, we assume that the world is composed of planar patches, and the obtained superpixels are their *one-to-many* 2D projection. This assumption represents a good estimate if the number of computed superpixels is large enough. We obtain the superpixels from an image by using an over-segmentation algorithm [6]. We represent the image as a set of superpixels $\mathbf{S}^t = \{S_1^t, S_2^t, ..., S_n^t\}$, where S_i^t defines superpixel i at time t. We define $\alpha_i^t \in \mathbb{R}^3$ the plane parameters associated to S_i^t such that for a given point $x \in \mathbb{R}^3$ on the plane satisfies $\alpha_i^t x = 1$. Our aim is to find the plane parameters for all superpixels in the image stream.

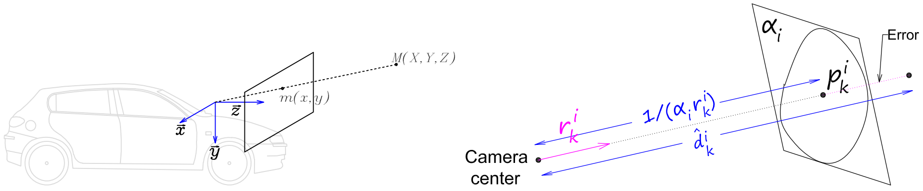
3.2 Spatial Depth Features

Spatial features for supervised depth estimation have not achieved much success compared to other computer vision domains such as object recognition and classification. Although the problem of monocular vision had been well studied in human vision (even before computers appear) and many monocular depth cues that human uses have been identified, however, it was not possible to obtain explicit depth representative measurements such as in stereo vision. Recently, there were several attempts to infer image 3D structure using spatial features and supervised learning [1][5][16]. In our method, we proceed in similar way, in order to capture texture information, the input image is filtered with a set of texture energies and gradient detectors (∽20 filters) [18]. Then by using superpixel segmentation image as a mask, we compute the filter response for each superpixel by summing its pixels in the filtered image. We refer the reader to [18] for more details. In order to capture general information, the aforementioned step is repeated for multiple scales of the image. Also, to add contextual information, e.g. texture variations, each superpixel feature vector includes the features of its neighbouring superpixels. Additionally, the formed feature vector includes colour, location, and shape features as they provide representative depth source for fixed camera configuration and urban environment. For instance, recognizing the sky and the ground. These features are computed as shown in table 1 in [14]. We donate X_i^t the feature vector for superpixel S_i^t.

3.3 Temporal Depth Features

In this subsection, we first describe some mathematical foundations and camera model. Then we explain how to perform sparse depth estimation which will be integrated in the probabilistic model given in subsection 3.5.

We use a monocular camera mounted on a moving vehicle. We assume that the Z axis of the camera coincides with the forward motion of the vehicle as shown in figure 1(a). Based on pin-hole camera model and camera coordinate system, a given $3D$ point $M(X, Y, Z)$ is projected on the $2D$ image as $m(x, y)$ by a perspective projection:



(a) (b)

Fig. 1. (a) Acquiring geometry: Camera installed on a moving vehicle with Z axis coincides with forward motion direction. (b) Illustration for how to compute the error in depth between the estimated value and the depth for a given α_i

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{f}{Z} \begin{bmatrix} X \\ Y \end{bmatrix} \tag{1}$$

When the vehicle moves, which is equivalent to fixed camera and moving world, the relationship between the velocity of a 3D point $[\dot{X}\dot{Y}\dot{Z}]^T$ and the velocity of its 2D projection $[\dot{x}\, \dot{y}]^T$ is given as the time derivative of equation 1. Then, based on the well-known optical flow equation $\dot{X} = -T - \Omega \times X$, and assuming a rigid scene, we decompose the 3D velocity into translational T and rotational velocity Ω [2]. Hence we obtain equation 2 which is the essence of most optical flow based SfM methods

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \frac{1}{Z} \begin{bmatrix} -f & 0 & x \\ 0 & -f & y \end{bmatrix} \cdot \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix} + \begin{bmatrix} xy/f & -f - (x^2/f) & -y \\ f + (y^2/f) & -xy/f & x \end{bmatrix} \begin{bmatrix} \Omega_x \\ \Omega_y \\ \Omega_z \end{bmatrix} \tag{2}$$

Based on this equation, we proceed in computing a sparse depth. We estimate the relative camera motion between two adjacent frames by first performing SIFT feature points matching [8]. Next we estimate the fundamental matrix using RANSAC [11] and bundle adjustment. Then, given camera intrinsic parameters, we can obtain the Essential matrix that encodes the rotation and translation between the two scenes. Which represent also the relative camera motion parameters $[T\ \Omega]$. The left hand side of equation 2 is basically the optical flow computed between two frames. In our implementation it is obtained using the well-known Lucas-Kanade with multi resolution and sub-pixel accuracy. Moreover, we benefit from the estimated Fundamental matrix to reject outliers in the optical flow. At this point, we could compute an approximate depth for the selected feature points.

Besides, given the specific camera setup as shown in figure 1(a), the motion of the camera is not totally free in the 3D space (motion of a vehicle). Therefore, we could add some constraints that express the feasible relative camera motion between two frames. For instance, limitation in T_y and Ω_z velocities. However, due to the absence of essential physical quantities, precise constraints on camera (or vehicle) motion could not be established theoretically. Instead, we evaluate experimentally possible camera transactions estimated from a set of video sequences acquired in different scenarios. As a result, we could establish some roles to spot outliers in the newly computed values for relative camera motion $[T\ \Omega]$. This way we improve the relative camera motion estimation in our case as we regularly have degenerated configurations (due to small baseline variations and dominant forward motion as mentioned earlier).

3.4 Occlusion Boundaries Estimation

When the camera translates, close objects move faster than far objects, and hence this causes to change the visibility of some objects in the scene. Although this phenomenon is considered as a problem in computer vision, it provides an important source of information about 3D scene structure. In our approach, we benefit from motion to infer occlusion boundaries. We use the method proposed in [7] to generate a soft occlusion boundaries map from two consecutive image frames. The method is based on supervised training of an occlusion detector thanks to a set of visual features selected by a Random Forest (RF) based model. Since occlusion boundaries lie close to surfaces edges, we use the classifier output as an indicator to the relationship between two superpixels if they are connected or occluded. Hence we add a penalty term in our MRF that forces the connectivity between superpixels. This term is inversely-proportional to the obtained occlusion indicator.

3.5 MRF for Depth Fusion

Markov Random Field (MRF) is becoming increasingly popular for modelling 3D world structure [1][5] due to its flexibility in terms of adding appearance constraints and contextual information. In our problem, we formulate our depth fusion as an MRF model that incorporates certain constraints with variable weights so they are jointly respected. Furthermore, we preserve the convexity of our problem such as in [1] to allow solving it through a linear program rather than probabilistic approaches for less computation time. We have seen earlier how to obtain temporal depth information, monocular features and occlusion boundaries. Figure 2 shows a simplified process flow for the proposed framework. We formulate our energy function which includes all of these terms as:

$$E(\alpha^t|X^t, O, \hat{D}, \alpha^{t-1}; \theta) = \underbrace{\sum_i \psi_i(\alpha_i^t)}_{\substack{\text{spatial depth}\\ \text{term}}} + \underbrace{\sum_{ij} \psi_{ij}(\alpha_i^t, \alpha_j^t)}_{\substack{\text{connectivity}\\ \text{term}}} + \underbrace{\sum_{ik} \phi_{ik}(\alpha_i^t, \hat{d}_k^i)}_{\substack{\text{temporal depth}\\ \text{term}}} + \underbrace{\sum_i \phi_i(\alpha_i^t, \alpha_i^{t-1})}_{\substack{\text{time consistency}\\ \text{term}}} \quad (3)$$

Where the superscripts t and $t-1$ refer to current and previous frames. X is the set of superpixels feature vectors. O is a map of occlusion boundaries computed from the frames t and $t-1$. The estimated sparse depth is $\hat{D}$, while $\hat{d}_k^i$ is the estimated depth value for pixel k in superpixel i. α_i is superpixel i plane's parameters and α is the set of parameters for all superpixels. θ are the learned monocular depth parameters. We now proceed in describing each term in our model (In the first three terms we will drop down the superscript of frame indicator t as they are the same).

Spatial Depth Term. This term is responsible for penalizing the difference between the computed plane parameters and the one estimated from spatial depth features (based on the learned parameters θ). It is given by the accumulated error for all pixels in the superpixel. See [18] P36-37 for details. For simplification, let's define a function $\delta(d_k^i, \hat{d}_k^i)$ that represents one point fractional depth error between an estimated value $\hat{d}_j^i$ and actual value d_j^i given plane parameters α_i. This potential function is given as

$$\psi_i(\alpha_i) = \beta_1 \sum_j \nu_k^i \delta(d_k^i, \hat{d}_k^i) \quad (4)$$

Where ν_k^i is a learned parameter that indicates the reliability of a feature vector X_k^i in estimating the depth for a given point p_k^i, see [1] for more details. β_1 is a weighting constant.

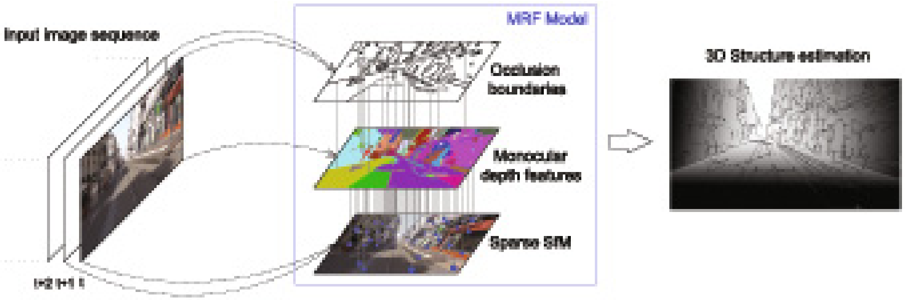


Fig. 2. Graphical representation of our MRF; it takes as input an image sequence. Occlusion boundaries and sparse SfM are estimated from two frames t and $t+1$, while monocular depth features are extracted from the current frame t, the MRF model integrate this information in order to produce a joint result for 3D structure estimation

Connectivity Prior. This term is based on the map of occlusion boundaries explained earlier. For each two adjacent superpixels, we compute an occlusion boundary indicator by summing up all pixels located at the common border in the estimated map. The obtained occlusion indicators are normalized so that they are in the range [0..1]. We refer o_{ij} for the indicator between superpixels i and j. The potential function is computed for each two neighbouring superpixels by choosing two adjacent pixels from each. The function penalizes the difference in distance between each of them to the camera. We have

$$\psi_{ij}(\alpha_i, \alpha_j) = \beta_2 \, o_{ij} \sum_{k=l=1}^{2} \delta(d_k^i, d_l^j) \tag{5}$$

Where β_2 is a weighting constant. This potential function forces neighbouring superpixels to be connected only if they are not occluded with the help of occlusion indicator o_{ij}. In comparison with the original method [1], we drop down the co-planarity constraint as we believe that the included temporal information and estimating occlusion boundaries indicator for motion provide an important source of depth information about plane orientation. Therefore, we do not mislead the estimation procedure with such approximation.

Temporal Depth Term. This term enforces some constraints that are established from the set of points where the depth is known. It is evident that with three non-collinear points we can obtain plane parameters α_i. However, to consider less or more number of points, we formulate this potential function to penalize the error between the estimated depth $\hat{d}_k^i$ for a point $p_k^i \in S_i$, and the computed depth given plane parameters α_i. Figure 1 (b) shows how this error is computed. Hence we have

$$\phi_{ik}(\alpha_i, \hat{d}_k^i) = \beta_3(\hat{d}_k^i - 1/{\alpha_i}^\top r_k^i) \tag{6}$$

Where β_3 is a weighting constant. we compute absolute depth error rather than fractional error since SfM is more confident than spatial depth estimation.

Time Consistency Term. In case of more than two frames, the quality of the 3D structure estimation varies from one frame to another, and it depends highly on the relative camera motion components (larger T_x and T_y translational motions results in better 3D structure estimation). Therefore we add some penalty in order to guide depth estimation at time t given the estimation at time $t-1$. This smooths the overall estimated structure variations in time. Hence, for each superpixel S_i^{t-1} we find its correspondence S_i^t based on the motion parameters and the size of common area. Additionally, we consider some visual features such as colour and texture. Eventually some superpixels will not have correspondence due to changing the field of view. We select the point p_k^i at the centre of the S_i^{t-1} and we form a ray from camera centre to this point. This ray intersects with superpixel S_i^t at point $p_k^{i\,\prime}$. The formulated potential function penalizes the distance across the ray between the two points

$$\phi_i(\alpha_i^t, \alpha_i^{t-1}) = \beta_4 \delta(d_k^{i\,\prime}, \hat{d}_k^i) \tag{7}$$

Here β_4 is smoothness term. We intend to use only one point to leave some freedom in plane orientation and for better 3D reconstruction refinement.

3.6 Parameters Learning and Inference

In our MRF formulation we preserve the convexity as all terms are linear or L_1 norm, which is solved using linear program. To learn the parameters, we first proceed with the first two terms of equation 3. We assume unity value for the parameters β_1 and β_2. The two parameters θ and ν are learned individually [18] using a dataset with ground-truth. For the rest of the parameters, β_1 and β_2 defines how the method is spatially oriented, while large β_3 turns the method into conventional SfM. β_4 allows previous estimation to influence the current one. Hence the weighting constants $\beta_{1..4}$ depends on the context, although they could be learned through cross-validation.

4 Discussion and Conclusion

We have presented a novel framework to perform 3D structure estimation from image sequence, which combines both spatial and temporal depth information to provide more reliable reconstruction. Temporal depth features are obtained using a sparse optical flow based structure from motion technique. The spatial depth features are obtained through a broad global and local feature extraction phase that tries to capture monocular depth cues. Both approaches have been tested independently on a wide set of images and proved to have good performance (see [19] [20] for comparison). This is why we believe that our joint approach gives better 3D structure estimation. Or at least, a performance similar to SfM technique (when the weight β_3 is assigned a large value). Additionally, it is adapted to our context where we regularly encounter a failure of certain depth source. For instance, in case of pure rotation in SfM or abnormal colors and appearance for some objects in spatial depth estimation.

References

1. Saxena, A., Sun, M., Ng, A.: Learning 3-d scene structure from a single still image. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8. IEEE (2007)
2. Aanæs, H.: Methods for structure from motion. IMM, Informatik og Matematisk Modellering, Danmarks Tekniske Universitet (2003)
3. Vedaldi, A., Guidi, G., Soatto, S.: Moving forward in structure from motion. In: IEEE Conference on CVPR 2007, pp. 1–7. IEEE (2007)
4. Saxena, A., Chung, S., Ng, A.: 3-d depth reconstruction from a single still image. International Journal of Computer Vision 76, 53–69 (2008)
5. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: IEEE Conference on CVPR, pp. 1253–1260. IEEE (2010)
6. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. International Journal of Computer Vision 59, 167–181 (2004)
7. Humayun, A., Mac Aodha, O., Brostow, G.: Learning to find occlusion regions. In: IEEE Conference on CVPR 2011, pp. 2161–2168. IEEE (2011)
8. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
9. Hartley, R., Zisserman, A., Ebrary, I.: Multiple view geometry in computer vision, vol. 2. Cambridge Univ. Press (2003)
10. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle Adjustment – A Modern Synthesis. In: Triggs, B., Zisserman, A., Szeliski, R. (eds.) ICCV-WS 1999. LNCS, vol. 1883, pp. 298–372. Springer, Heidelberg (2000)
11. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24, 381–395 (1981)
12. Lindeberg, T., Garding, J.: Shape from texture from a multi-scale perspective. In: Fourth International Conference on Computer Vision, pp. 683–691. IEEE (1993)
13. Torralba, A., Oliva, A.: Depth estimation from image structure. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 1226–1238 (2002)
14. Hoiem, D., Efros, A., Hebert, M.: Automatic photo pop-up. ACM Transactions on Graphics 24, 577–584 (2005)
15. Hoiem, D., Efros, A., Hebert, M.: Recovering surface layout from an image. International Journal of Computer Vision 75, 151–172 (2007)
16. Sturgess, P., Alahari, K., Ladicky, L., Torr, P.: Combining appearance and structure from motion features for road scene understanding (2009)
17. Bao, S., Savarese, S.: Semantic structure from motion. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2025–2032. IEEE (2011)
18. Saxena, A.: Monocular depth perception and robotic grasping of novel objects. Stanford University (2009)
19. Saxena, A.: State-of-the-art results of the depth prediction from single image. Website (2012), http://make3d.cs.cornell.edu/results_stateoftheart.html
20. Civera, J., Davison, A., Montiel, J.: Structure from Motion Using the Extended Kalman Filter, vol. 75. Springer (2011)