Uncalibrated Camera Based Interactive 3DTV

M.S. Venkatesh, Santanu Chaudhury, and Brejesh Lall

Department of Electrical Engineering, IIT Delhi, India
msvenka@gmail.com, santanuc@ee.iitd.ac.in, brejesh@ee.iitd.ac.in

Abstract. In this paper we propose a novel architecture for an interactive 3DTV system based on multiple uncalibrated cameras placed at general positions. The signal representation scheme proposed is compatible with the standard multi view coding framework making it amenable to using existing coding and compression algorithms. The proposed scheme also fits naturally to the concept of *true* 3DTV viewing experience where the viewer can choose a novel viewpoint based on the contents of the scene.

Keywords: Interactive 3DTV, Uncalibrated cameras, Plane sweeping, Depth map.

1 Introduction

In this paper we propose a novel architecture for interactive 3DTV which enables a viewer to choose a 3D view of the scene based on scene content. We assume that the scene is captured by multiple uncalibrated cameras in general positions. We use a statistical learning method to automatically detect scene constraints that are then used to calibrate the cameras as well as provide orientation information to generate depth maps for each view. The advantage of our signal representation scheme is that it is compatible with the standard Multi View Coding framework of video plus depth[3,4]. This enables the direct application of the existing coding and compression algorithms. The scene orientation information can be easily embedded in the meta-data part of the signal. The novelty of our architecture is that the scene orientation information allows the user to select different viewpoints based on the scene content to interactively generate novel 3D views at the receiver's end. Typical applications of our methodology is in 3D viewing of outdoor scenes like monuments, urban buildings etc. The user can interactively specify which part of the monument he wants to view in 3D. The typical characteristics of such scenes like orthogonal planes assist in automatic detection of scene constraints and orientations.

3DTV is considered as the next step in enhancing the user viewing experience. A *true* 3DTV is supposed to give a 3D view of the scene as well allow the user to choose different viewpoints of the scene. Our approach fits naturally in this context. The video plus depth format [3] in which each video is accompanied by a per pixel depth data is considered most suitable for generation of high quality 3D views as well as intermediate view synthesis at the receiver. To generate

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 213–219, 2011.

the depth map, camera calibration information is required. Camera calibration from scene constraints and generation of depth maps from stereo and multi view images are topics that have been extensively studied in computer vision [15]. In our methodology we automatically discover these scene constraints and exploit them for the calibration of the cameras. Plane sweeping algorithms like [5,6,7] outperform standard stereo matching algorithms for generation of depth map in presence of oblique structures in scenes. The directions along which the planes are swept are determined by analysis of a sparse or dense set of 3D points. In our approach, the scene classification step itself gives us the orientations used for plane sweeping. The dense depth map thus generated is then used to generate existing as well as interpolate novel stereo views of the scene without performing an explicit 3D reconstruction using a Depth Image Based Rendering technique [4]. These novel views are specified by new camera viewpoints with respect to the existing viewpoints.

Some work has been done regarding user interaction in 3DTV [12]. The authors in [13] allow for view switching, frozen moment and view sweeping at the receiver end. The novelty of our approach is that it allows the user to choose the viewpoint based on scene content to be able to view novel 3D views of parts of the scene as well as arbitrary novel views of the complete scene. Automatically extracting the scene constraints and orientations from the images to use them for camera calibration is also a novelty of our work. Another novelty of our method is that we obtain the direction of family of planes and the confidence measures from the scene classification [1] step itself.

In section 2, we present the basic outline of our system architecture. Section 3 outlines the novel view viewpoint selection and the DIBR technique used to synthesize novel views. In section 4 we present experimental results and discussions followed by the conclusion and future work.

2 System Architecture

We illustrate our proposed methodology with an example of a building as shown in the figure 1. Images are taken from multiple uncalibrated cameras placed at general positions. Then classifier of [1] is used to classify the image and identify the horizontal(green) and vertical (red) planes. This constraint is then used to calibrated the cameras. The orientations of scene (indicated by arrows) are also recovered along with their confidence measures. This is used to generate a depth map. The orientation information along with the video plus depth forms the signal representation. At the receiver end the user interactively chooses certain parts of the scene to view and then the orientation information in that part of the image is used to obtain the novel 3D views of the scene.

2.1 Scene Classification

The authors in [1] classify the outdoor scene into geometric classes that depend on the orientation of the particular object in the scene based on statistical

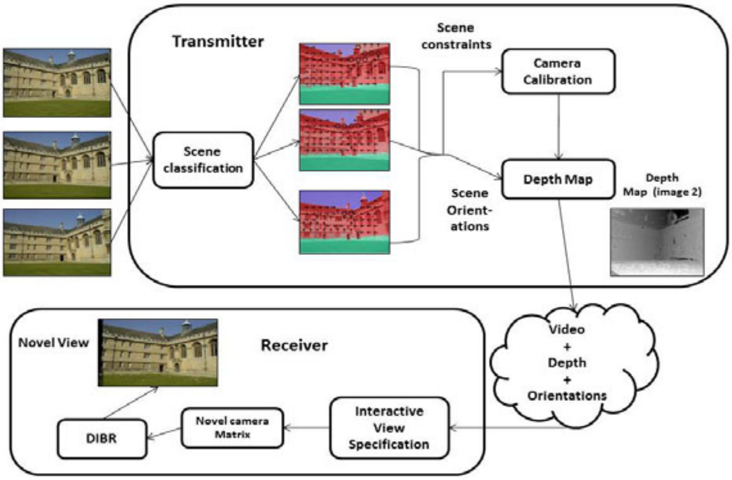


Fig. 1. Outline of the proposed architecture

learning. They classify an image into planar vertical regions (facing left, center, or right), ground (horizontal) regions, sky and non planar regions (porous and solid) without using any calibration information. A confidence measure is also associated with each class.

2.2 Camera Calibration

Many methods exist for calibrating the cameras using the scene constraints. Specifically, we use the orthogonality of the vertical and horizontal planes obtained from the scene classification step for calibration. We obtain automatic point correspondences from SIFT features [1] and filter mismatches by using epipolar constraint [2]. Using these point correspondences, we obtain a projective reconstruction and then impose the constraints of the absolute quadric as well as the scene constraints to upgrade it to a Euclidean one [15,2].

2.3 Obtaining the Dense Depth Map

We follow an approach closely related to [5]. Here every pixel is assigned a label which indicates which plane it belongs to and an energy function which penalizes abrupt changes in surface normals is minimized [8,9,10,11]. The confidence measures obtained in the scene classification step is incorporated into the cost function.

[1] http://www.vlfeat.org/

[2] http://www.csse.uwa.edu.au/~pk/research/matlabfns/

The normal to the planes (vertical planes and ground plane) are used to define "k" families of parallel planes Π_{ki} given by $n_k^T X + d_{ki} = 0$ where $d_{ki} \in [d_{min}, d_{max}]_k$ The range $[d_{min}, d_{max}]$ for each of the plane families are obtained empirically. For each of the planes we find the 3x3 homography induced between two images r and s [15].

This homography is used to warp the other images onto an image I_r for whom depth map is to be generated. Then for warped image I_w, we obtain a cost for every pixel (x, y) in the image I_r, for each of the planes Π_{ki} as

$$cost(x, y, \Pi_{ki}) = \sum_{(p,q) \in N} |I(x-p, y-q) - I_w(x-p, y-q)| - w * log(F(\Pi_{ki}))$$

where w is a weight factor determined by experiments and $F(\Pi_{ki})$ is the probability that the pixel belongs to the ki^{th} plane obtained from the confidence measures generated after the scene classification step. N is defined neighbourhood of the pixel taken to be 3x3.

We formulate an energy function similar to [5] and we find out the depth $Z(x, y)$ of each pixel (x, y) of image I_r. The depth map thus obtained is used in virtual view synthesis which is explained in the next section.

3 Virtual View Synthesis

A new virtual camera is defined by its camera parameters R_v, K_v, and C_v [15]. Given an image I_r whose camera parameters are R_r, K_r and C_r and its corresponding depth map Z_r, The image I_v is generated by [4]

The resulting virtual views have holes which are the results of either erroneous depth values or due to truncation of the pixel positions. To fill these holes, the homography corresponding to the nearest non zero plane label is computed and using it, the pixel value is transferred from the given image.

3.1 Interactively Choosing the Viewpoint

The user interactively selects a part of the scene through an interface like a mouse pointer or remote. For instance if a wall is selected, the virtual camera has to be placed such the wall is fronto parallel. The orientation of this part of the image obtained after scene classification gives the direction in which the camera has to be rotated. Once the direction is determined, the following steps are followed.

1. A rotation matrix R' is obtained by choosing an angle θ bounded by the angle between the normal to the plane (wall) and normal to the principal plane of the camera and a small translation t' is chosen such that the virtual camera undergoes a little translation so as to keep the view within the image bounds. This Euclidean transformation is used to get the destination camera centre C_v and rotation matrix R_v.

2. K_v is chosen to be the same as K_r. The destination camera matrix $P_v = K_v R_v [I| - C_v]$
3. The new camera matrices are chosen by interpolation between the camera P_r and P_v with interpolation parameter $\lambda \in [0..1]$ as
$C_\lambda = (1 - \lambda)C_r + \lambda C_v$
$R_\lambda = slerp(R_r, R_v, \lambda)$
$P_\lambda = K_\lambda R_\lambda [I| - C_\lambda]$
where *slerp* refers to spherical linear interpolation.
4. To generate stereo pairs, the shift sensor algorithm [4] is used.

4 Results and Discussions

We illustrate the various steps of our proposed architecture on two standard data sets Wadham College and Merton College.[3] For Wadham College data set, we generate intermediate image sequences when the user wants to look at the left wall. For the Merton data set we generate sequences when the user wants to look at the right wall. Both of these sequences are generated from the image no 2 of each data set. As the rotation increases, the rendering quality decreases. This is because of erroneous depth values and also due to mis-classification.

The average 2D re projection errors (in pixel) from the camera calibration are shown below

Data Set	camera 1	camera 2	camera 3
Merton College	0.410790	0.343800	0.311718
Wadham College	0.405233	0.398542	0.295823

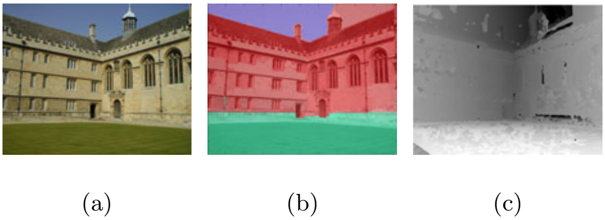

(a) (b) (c)

Fig. 2. Fig (a) shows image 2 of Wadham sequence. Fig (b) shows its scene classification and Fig (c) is the depth map extracted.

The stereo views were generated and were displayed on a samsung 3D LED TV. Subjective rendering quality measures were assigned based on a scale of 1 to 5 in increasing order of quality. The subjective scores are shown below. The sequences refer to the intermediate stereo views generated.

[3] `http://www.robots.ox.ac.uk/~vgg/data2.html`

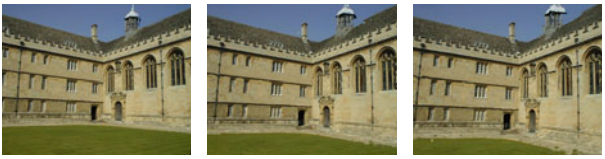
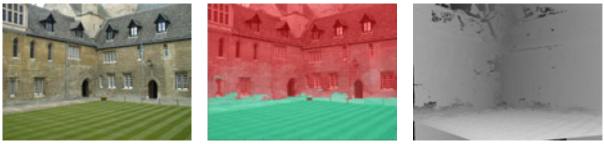


Fig. 3. Some of the synthesized image sequences of Wadham when the user wants to "look" at the left wall



(a) (b) (c)

Fig. 4. Fig (a) shows image 2 of Merton sequence. Fig (b) shows its scene classification and Fig (c) is the depth map extracted.

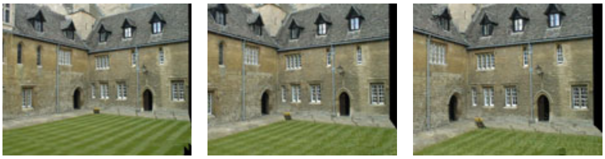


Fig. 5. Some of the synthesized image sequence of Merton when the user wants to "look" at the right wall

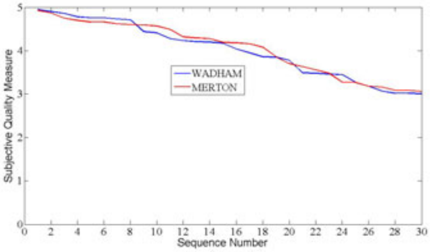


Fig. 6. Subjective Quality

5 Future Work and Conclusion

In this paper we have proposed a novel architecture for interactive rendering of outdoor scenes for 3DTV. The interactive selection of viewpoints based on scene content offers a different approach to novel view synthesis. We have worked with images. The extension to videos is straight forward. As part of the future work , we would also like to investigate the possibility of detecting objects in the scene and allow the user to view these objects interactively. Another extension would be in switching between groups of cameras and allowing the user to view objects captured by different groups.

Acknowledgement

This work was in part supported by the Department of Science and Technology, Govt. of India. (SR/S3/EECE/0006/2010). We would also like to thank Ayesha Chaudhary for valuable feedback.

References

1. Hoeim, D., Efros, A.A., Hebert, M.: Geometrical context from a single image. In: ICCV (2005)
2. Svoboda, T., Martinec, D., Pajdla, T.: A convenient multi-camera self-calibration for virtual environments. PRESENCE: Teleoperators and Virtual Environments 14(4), 407–422 (2005)
3. Muller, K.: View Synthesis for Advanced 3D Video Systems. EURASIP Journal on Image and Video Processing 2008, article ID 438148, 11 pages (2008), doi:10.1155/2008/438148
4. Fehn, C.: A 3D-TV Approach Using Depth-Image-Based Rendering (DIBR). In: Proceedings of 3rd IASTED Conference on Visualization, Imaging, and Image Processing, September 2003, pp. 482–487 (2003)
5. Gallup, D., Frahm, J.M., Mordohai, P., Yang, Q., Pollefeys, M.: Real-Time Plane-Sweeping Stereo with Multiple Sweeping Directions. In: CVPR (2007)
6. Micusik, B., Kosecka, J.: Multi-view Superpixel Stereo in Urban Environments. International Journal of Computer Vision 89(1) (2010)
7. Sinha, S.N., Steedly, D., Szeliski, R.: Piecewise Planar Stereo for Image-based Rendering. In: ICCV (2009)
8. Boykov, Y., Veksler, O., Zabih, R.: Efficient Approximate Energy Minimization via Graph Cuts. IEEE TPAMI 23(11) (November 2001)
9. Kolmogorov, V., Zabih, R.: What Energy Functions can be Minimized via Graph Cuts? IEEE TPAMI 26(2), 147–159 (2004)
10. Boykov, Y., Kolmogorov, V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. IEEE TPAMI 26(9), 1124–1137 (2004)
11. Delong, A., Osokin, A., Isack, H.N., Boykov, Y.: Fast Approximate Energy Minimization with Label Costs. In: CVPR (2010)
12. Fehn, C., De La Barr, R., Pastoor, S.: Interactive 3-DTV.Concepts and Key Technologies. Proceedings of IEEE 94(3) (March 2006)
13. Lou, J., Cai, H., Li, J.: A RealTime Interactive MultiView Video System. In: MM 2005 Proceedings of the 13th ACM conference on Multimedia (2005)
14. Kubota, A., Smolic, A., Magnor, M., Tanimoto, M., Chen, T., Zhang, C.: Multiview Imaging and 3DTV. IEEE Signal Process. Mag. 24(6), 10–21 (2007)
15. Hartley, R., Zisserman, A.: Multiple view geometry in Computer Vision, March 2004. Cambridge University Press, Cambridge (2004)