Fixed Point Learning Based 3D Conversion of 2D Videos

Nidhi Chahal(✉) and Santanu Chaudhury

Indian Institute of Technology, Delhi, India
{nidhi.vce,schaudhury}@gmail.com

Abstract. The depth cues which are also called monocular cues from single still image are more versatile while depth cues of multiple images gives more accurate depth extraction. Machine learning is a promising and new research direction for this type of conversion in today scenario. In our paper, a fast automatic 2D to 3D conversion technique is proposed which utilizes a fixed point learning framework for the accurate estimation of depth maps of query images using model trained from a training database of 2D color and depth images. The depth maps obtained from monocular and motion depth cues of input images/video and ground truth depths are used in training database for the fixed point iteration. The results produces with fixed point model are more accurate and reliable than MRF fusion of both types of depth cues. The stereo pairs are generated then using input video frames and their corresponding depth maps obtained from fixed point learning framework. These stereo pairs are put together to get the final 3D video which can be displayed on any 3DTV and seen using 3D glasses.

Keywords: Monocular · ZPS · Artifacts

1 Introduction

The world of 3D includes the third dimension of depth which can be perceived in form of binocular disparity by human vision. The different views of real world are perceived by both eyes of human as these are located at different positions. From these different views, the brain is able to reconstruct the depth information. A 3DTV display presents two slightly different images of every scene to the individual eyes and thus 3D perception can be realized.

There is great impact of depth map quality on overall 3D output. The better visualization of 3D views demands accurate and denser depth map estimation. In general, depth cues of multiple images gives more accurate depth extraction. And monocular depth cues which are utilized for depth estimation from single still image are more versatile. For the accurate and high quality conversion of 2D images to 3D models, a single solution does not exist. To enhance the accuracy of results, various depth cues should be combined in such a way that more dense and fine depth maps can be obtained. The depth consistency should also

M. Kryszkiewicz et al. (Eds.): PReMI 2015, LNCS 9124, pp. 85–94, 2015.
DOI: 10.1007/978-3-319-19941-2_9

be maintained for the accurate depth estimation. The principle of depth consistency states that if the color values or intensities of the neighboring pixels are similar, these also should have similar depth values. In our paper, machine learning framework is introduced which gives faster and more accurate 2D to 3D conversion. In training phase, the video frames are extracted first from training video and their appearance features are utilized; the depth values of neighboring blocks of a given image block are used as contextual features. Thus, a trained model is obtained which is used in testing phase where new testing video frames are used as input and their depth values are obtained as output using this learning framework. So, a 3D structure of testing video frames is obtained from a training database of video frames and their depths.

The depth extraction is achieved by using monocular depth cue of images/video frames considering single image at a time and motion depth cue using more than one image. The contextual prediction function is used in this model which gives labeling i.e. assigning depth values of image blocks or individual pixels as output while input being both its features and depth values of rest of the image blocks or pixels. Finally, stereo pair generation is achieved by using input images and their corresponding depth maps using image warping technique. And thus, final 3D output images/video is obtained which can be displayed on 3DTV display and 3D output can be viewed using 3D glasses.

2 Related Work

In the last decade, a new family of machine learning approaches are introduced for the depth estimation of different 2D images. In these techniques, training database of color and depth images are used to estimate the depth map of a query color image. In this way, the information is transferred from the structure correlations of the color and depth images in the database to the query color image and its 3D structure is obtained. Some 3D conversion systems have used monocular depth cues only for the depth extraction from single still image which has their own limitations. Most of the prior work have focused on obtaining the better quality of estimated depth maps, but at the expense of high computational cost algorithms.

A depth estimated in [1] from a single color image using MRF framework and image parsing processing. The machine learning approach has also been adopted in [2] that exchanges transference of labels by directly depth map data. The semantic labels and a higher complex supervised model is incorporated in [3] to achieve more accurate depth maps. Konard in [4] has not used image registration step to reduce the computational burden of previous approaches and used a matching based search framework based on HOG features to find similar structured images. This approach has less computational cost than previous approaches, but for many practical applications, it is still too high. The depth maps are combined using MRF model which incorporates contextual constraints into fusion model described in [7], but depth learning is not done in this method. MRF and CRF are often limited to capturing less neighborhood interactions due to heavy computational burden in training and testing stages and thus, their modeling capabilities are also limited.

In our paper, fixed point learning framework provides depth learning from training database of color images, local depth values from focus and motion cues and their ground truth depth values. The images features and depth values from focus and motion cues are used as training data. And we are using ground truth depth values which we are obtained after manual labeling of input images because ground truth depth maps are available for limited data sets. These ground truth depths have been used as training labels which results to fixed point iteration. The learned fixed point function captures rich contextual/structural information and is easy to train and test and much faster also which balances the performance and learning time. For the testing images, reliable depth maps can be obtained using this trained model. The proposed learning framework provides automatic and fast 2D to 3D conversion method which converges at very less number of iterations with good accuracy in results. Fixed point model [8] is a simple and effective solution to the structured labeling problem. We compared the results with MRF fusion as described in [7] of depth cues. Depth maps from focus and motion are fused using MRF approach. But the depths obtained from fixed point model have higher accuracy than MRF fused depths. The results are also obtained with other fusion methods like weighted averaging, least squares, maximizing approach, window technique. All the fusion depths have less accuracy as compared to fixed point model results.

3 Depth Estimation

The HVS (Human Visual System) exploits a set of visual depth cues to perceive 3D scenes. The extraction of depth information from single view is known as monocular cue. There are various monocular depth cues as shading, texture gradient, accommodation, linear perspective and others. The motion depth cue provides depth extraction from multiple images. Both methods are used in the paper to utilize the benefits of both depth cues. The two important depth cues for depth extraction are focus/defocus and motion parallax on the basis of their relative importance in the human brain. The importance depends on the distance (JND-Just noticeable difference) from the observer. A depth cue with larger JND means that it is harder to be detected in the human vision system.

3.1 Depth Extraction from Monocular Depth Cue

The defocusing factor is used as monocular depth cue as discussed above. The principle that stuff which is more blurry is further away is used in depth estimation from focus depth cue. The method used here recovers depth from a single defocused image captured by uncalibrated conventional camera as described in [5]. From the blur amount at each edge location, a sparse depth map is estimated. Then, propagate the depth estimates to the whole image and obtain full depth map. The diameter of CoC characterizes the amount of defocus and can be written as

$$c = \frac{|d - d_f|}{d} \cdot \frac{f_0^2}{N(d_f - f_0)} \tag{1}$$

where d_f is the focus distance, d is the object distance f_0 and N are the focal length and the stop number of the camera respectively.

3.2 Depth Extraction Using Motion Parallax

The other depth cue used is depth from motion. There are two ways to calculate depth from motion parallax and these are motion blur and optical flow. The drawback of depth extraction using motion blur is the cost due to the multiple of high quality cameras, image processing programming and a big hardware background. In our experiments, optical flow is used for depth extraction and it is calculated between two consecutive frames taken from the scene as described in [6]. The length of the optical flow vectors will be inverse proportional to the distance of the projected point. The optical flow can also be used to recover depth from motion by using:

$$Z = v_c \cdot \frac{D}{V} \tag{2}$$

where v_c is the velocity of camera,
D is the distance of the point on image plane from focus of expansion,
V is the amplitude of flow and
Z is the depth of the point in the scene projected.

4 Fixed Point Learning Based 2D to 3D Conversion

There are infinite number of possible solutions for recovering 3D geometry from single 2D projection. The problem can be solved using learning based methods in vision. Initially a set of images and their corresponding depth maps are gathered in supervised learning. Then suitable features are extracted from the images. Based on the features and the ground truth depth maps, learning is done using learning algorithms. The depths of new images are predicted from this learned algorithm. The local as well as global features are used in a supervised learning algorithm which predicts depth map as a function of image. If we take examples of monocular depth cues, local information such as variation in texture and color of a patch can give some information about its depth, these are insufficient to determine depth accurately and thus global properties have to be used. For example, just by looking at a blue patch, it is difficult to tell whether this patch is of a sky or a part of a blue object. Due to these difficulties, one needs to look at both the local and global properties of an image to determine depth.

The entire image is initially divided into small rectangular patches which are arranged in a uniform grid. And a single depth value for each patch is estimated. Absolute depth features are used to determine absolute depth of a patch which captures local feature processing. To capture additional global features, the features used to predict the depth of a particular patch are computed from that patch as well as the neighboring patches which is repeated at each of the multiple

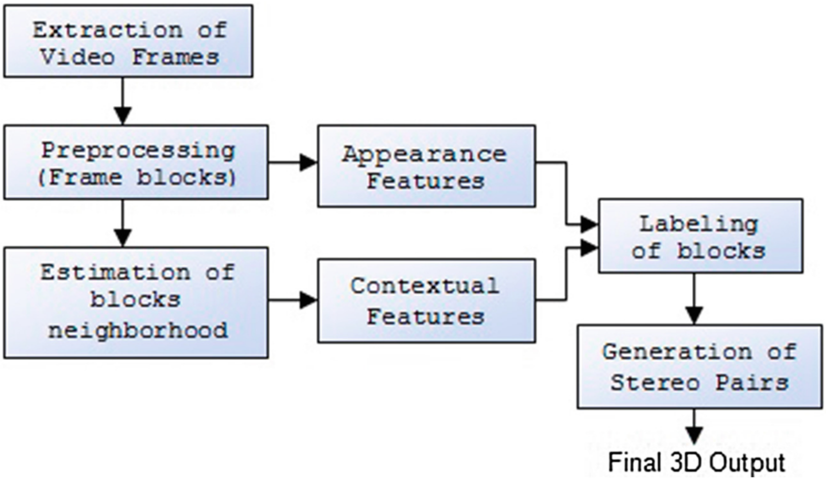

Extraction of Video Frames
Preprocessing (Frame blocks)
Appearance Features
Estimation of blocks neighborhood
Contextual Features
Labeling of blocks
Generation of Stereo Pairs
Final 3D Output

Fig. 1. 2D video to 3D output block diagram

scales, so that the feature vector of a patch includes features of its immediate neighbors at large spatial scale, its far neighbors at a larger spatial scale, and again its very far neighbors at an even larger spatial scale. The fixed point model approach is an effective and simple solution to the problem of structured labeling in image processing. A fixed point function is obtained with the labeling of the structure being both the output and the input. The overall fixed point function is a vector form of contextual prediction function of the nodes. The contextual prediction function gives labels of an individual node as output while inputs are both its features and labeling of rest of the nodes or neighboring nodes. The learned function captures rich contextual/structural information and is easy to train and test.

In image processing, a structured input is an image of all pixels and the structured outputs are the corresponding labels of these pixels. The structured input as shown in Fig. 3a, can be represented as a graph, denoted as $\chi = (\mathbf{I}, \mathbf{J})$. The block $b_i \in \mathbf{I}$ represents i^{th} block with its features denoted as f_i. And $\mathbf{J}$ represents the neighborhood of each block. M denotes the number of neighbors a block can have in its neighborhood specification. That is, M number of blocks specifies the neighborhood N_i of block b_i in four directions: left, right, top and bottom. The labeling of the neighborhood of b_i is denoted as $\mathbf{q}\ N_i$ where $\mathbf{q}$ denotes labeling of all the blocks in χ. The labels for all blocks can be represented as $\mathbf{p} = (p_i : 1...\phi)$, where $p_i \in \phi$ and ϕ is the label space. For each block b_i, a context prediction function Ψ takes both b_i's appearance features f_i and contextual features $\mathbf{q}N_i$ as inputs. And this prediction function predicts the labeling $p_i \in \phi$ for the blocks of image. The contextual prediction function defined in [8] can be represented as follows:

$$q_i = \psi(f_i, \mathbf{q}N_i; \delta) \tag{3}$$

where ψ is a regression function within range [0,1] and δ is the parameter of the function. The labeling $\mathbf{q}$ of all image blocks can be written as

$$\mathbf{q} = \psi(f_1, f_2, ...f_n, \mathbf{q}; \delta) \tag{4}$$

Given the depth values $\mathbf{q}$ and features f_1, f_2,...f_n of training data, the parameter δ is learned. In experiments, SVM is used as contextual prediction function. The function is trained for each image block on the basis of appearance features f_i and contextual features $\mathbf{q}N_i$. The contraction mapping, when learned is applied iteratively to the new structured inputs. For new structured input, the label q_i of a block $b_i \in \mathbf{I}$ is initialized with a value. It is taken as zero during implementation as it does not effect the accuracy of results.

4.1 Pre Processing and Appearance Features

The frames are extracted first from the 2D input video and converted to gray scale. These frames are divided into 8×8 blocks to reduce the computations. The appearance features of these input frames/images is extracted then. The examples of appearance features of image are maximum height of the block, width of the block, aspect ratio and mean wise averaging of the image blocks. The RGB values can also be taken as appearance features of images.

4.2 Neighborhood and Contextual Features

The neighborhood of each image block is identified then. The parameter M defines the span of neighborhood of image blocks. The neighboring blocks of i^{th} block is defined by all the adjacent blocks to its left, right, top and bottom. In our experiments, we used M = 1 to 6 where M = 6 gives the best results. A normalized histogram of neighboring blocks is used as contextual feature $\mathbf{q}N_i$. For each block b_i, a 4 × M × C dimensional contextual feature vector is created. In this vector, 4 specifies the neighborhood in four directions (upper, lower, left and right), M is the span of the context and C is the number of class labels. In implementations, following values are used.

M = 6 = Number of neighbors a node can have in the neighborhood specification.
C = 255 = No of classes = No of histogram bins
Number of iterations in prediction = 3
Constant for labeling vector = 0.

4.3 Labeling of Blocks

All the image blocks are labeled in this step. Labeling means assigning depth values to each block of an image. The ground truth depth values are used as training labels resulting in fixed point iteration. At each iteration, the depth value for which convergence is achieved, is assigned to the corresponding block during testing process. The fixed-point model converges very quickly at the testing stage with only 2 to 3 iterations. We have used three number of iterations in our experiments as it gives good accuracy with less implementation time. We experimented with L1 regularized support vector machine (SVM-L1), provided in the Lib linear software package as the classifier. The SVM parameter is taken as '−s 5 −c 1' and the constant of labeling vector is taken as 0.

5 Stereo Generation

The stereo pairs are generated at the end from 2D original images and final obtained depth maps. Consider original 2D image as intermediate image and generate left and right views from this image. The right and the left images will be half shifted or warped toward the respective direction and all the artifacts introduced by the processing will be part of both images but halved visible. So, less artifacts are produced in this method. We applied this technique [9] for final stereo generation.

First, pre-processing of depth maps is done by shifting depth map by ZPS i.e. Zero Parallax Setting:

$$Z_c = \frac{Z_n - Z_f}{2} \tag{5}$$

where Z_n and Z_f are the nearest clipping plane and the farthest clipping plane of the depth map. In an 8-bit depth map, $Z_n = 255$ and $Z_f = 0$. After that, the depth map is further normalized with the factor of 255, so that the values of the depth map lie in the interval of $[-0.5, 0.5]$, values that are required by the image warping algorithm. Then, Gaussian filter is used for depth smoothing.

w = 3 $\times \sigma$ where w is the filter window size
σ is standard deviation
Depth smoothing strength = Baseline/4

The stereo pair is obtained by transferring the original points at location (P_m,y) to left points (P_l,y) and right points (P_r,y):

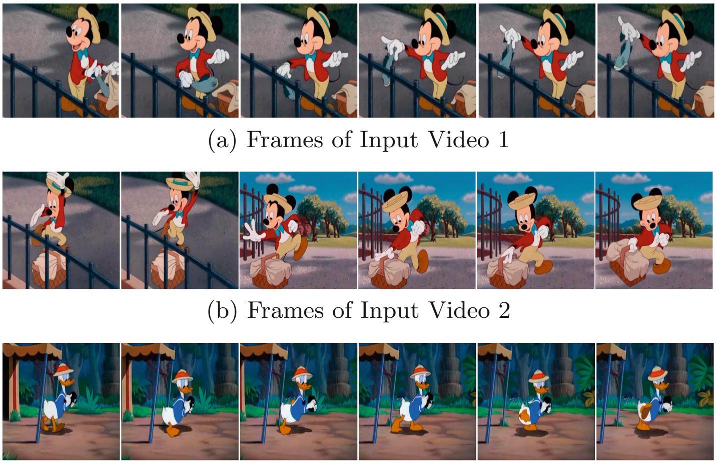
$$P_l = P_m + \frac{Bf}{2Z} \tag{6}$$

$$P_r = P_m - \frac{Bf}{2Z} \tag{7}$$

where B is the baseline distance between two virtual cameras.
B = 5 percent of width of depth image
f is the focal length of camera
f is chosen as one in our experiments without any loss of generality.

6 Implementation Results

The input color video of Mickey mouse is taken as input here as shown in Fig. 2a. The video is in wmv format and the resolution of video frames is 960 × 1080.



(a) Frames of Input Video 1

(b) Frames of Input Video 2

(c) Frames of Input Video 3

Fig. 2. Examples of 2D video frames to be converted to 3D video

The resolution of final 3D image is 1920 × 1080 where left and right frames are put side by side. The results of fixed point model for some video frames in form of accuracy (percentage) is shown in Table 1 as follows:

Table 1. Fixed point implementation results

Image	Accuracy with fixed point model	Accuracy with MRF fusion
1	83.57	73.6
2	83.45	73.72
3	82.71	74.90
4	82.39	74.00
5	83.00	74.72
6	82.38	73.73

Here, results are shown in Fig. 3 where (a) and (b) shows original input frames of input video in RGB format. Figure 3c shows depth from focus, (d) shows depth from motion (optical flow) and (e) shows the final 3D output which can be shown on 3DTV display using 3D glasses. The examples of other cartoon and human video frames are also shown in Fig. 4. The subjective test is conducted to show the quality of 3D images, that is, left and right image pair for every 3D image on 3DTV display by wearing 3D glasses. Five participants involved in the experiment to evaluate 3D quality. Their individual scores are mentioned in Table 2 using which MOS i.e. mean opinion score is calculated. The MOS is the arithmetic mean of all the individual scores and can range from 1 (worst) to 5 (best).

Table 2. Subjective quality measure

Video	Viewers score	MOS
Mickey	4,4,5,4,4	4.2
Donald	4,4,4,4,5	4.2
Human	4,3,5,4,4	4

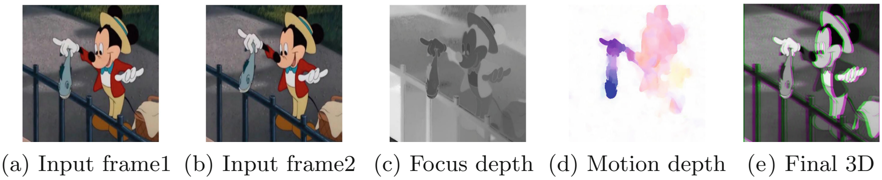
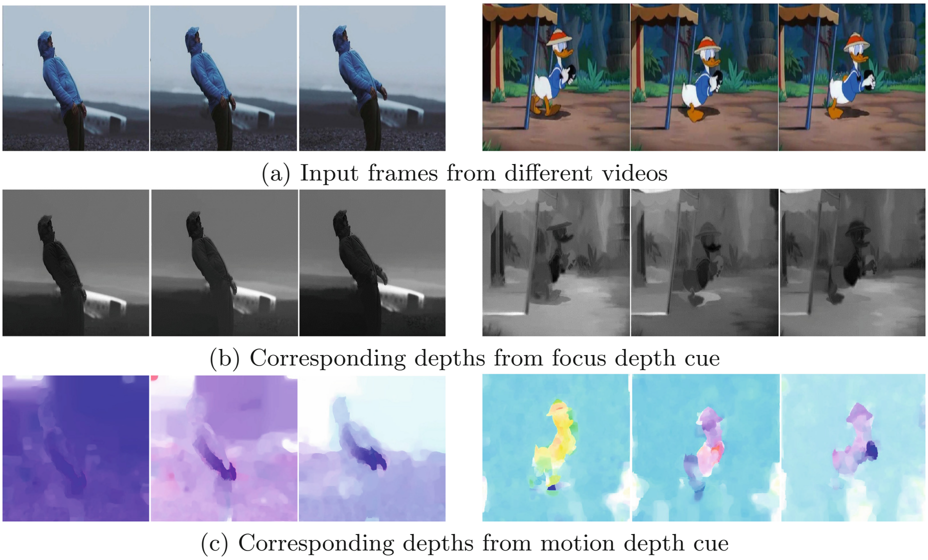

(a) Input frame1 (b) Input frame2 (c) Focus depth (d) Motion depth (e) Final 3D

Fig. 3. Example of two 2D frames to final 3D output


(a) Input frames from different videos

(b) Corresponding depths from focus depth cue

(c) Corresponding depths from motion depth cue

Fig. 4. Examples of other video frames

7 Conclusion

The depth estimation is the key step for converting 2D to 3D images or videos. For high quality and accurate 3D output, the depth maps should be reliable. In this paper, depth extraction has been done using depth cue of single still image and motion cues also which use more than one image. Fixed point model uses ground truth depth maps and provides learning framework for accurate estimation of depths which gives higher accuracy than MRF and other fusion methods. Finally, stereo pair generation is done to get the final 3D output which can be viewed on any 3DTV display. The quality of 3D images is also evaluated using MOS which indicates good and reliable depth extraction. MRF has limitations in capturing neighborhood interactions which limits their modeling capabilities while fixed point learned function captures rich contextual/structural information and is easy to train and test. The model is much faster to train and balances the performance and learning time giving higher accuracy than MRF depth fusion.

References

1. Saxena, A., Sun, M., Ng, A.Y.: Make3d: learning 3d scene structure from a single still image. Trans. Pattern Anal. Mach. Intell. **31**, 824–840 (2009). IEEE
2. Konrad, J., Wang, M., Ishwar, P.: 2d-to-3d image conversion by learning depth from examples. In: Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 16–22. IEEE (2012)
3. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing: label transfer via dense scene alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1972–1979. IEEE (2009)
4. Konrad, J., Wang, M., Ishwar, P., Wu, C., Mukherjee, D.: Learning-based, automatic 2d-to-3d image and video conversion. Trans. Image Process. **22**(9), 3485–3496 (2013). IEEE
5. Zhuo, S., Sim, T.: Defocus map estimation from a single defocused image. Pattern Recogn. **44**(9), 1852–1858 (2011)
6. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2432–2439 (2010)
7. Xu, M., Chen, H., Varshney, P.K.: An image fusion approach based on markov random fields. Trans. Geosci. Remote Sens. **49**(12), 5116–5127 (2011)
8. Li, Q., Wang, J., Tu, Z.: Fixed-point model for structured labeling. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, vol. 28 (2013)
9. Kang, Y., Lai, Y., Chen, Y.: An effective hybrid depth-generation algorithm for 2D-to-3D conversion in 3D displays. J. Disp. Technol. **9**(3), 154–161 (2013)