Main vertical objects detecting algorithm for real-time 2D-to-3D conversion

Chunyu lin, Jan De Cock, Yao Zhao [2], Peter Lambert and Rik Van de Walle
Multimedia Lab-Ghent University-IBBT
Gaston Crommenlaan 8 box 201 9050 Ledeberg-Ghent, Belgium
Institute of Information Science, Beijing Jiaotong University [2]
Beijing Key Laboratory of Advanced Information Science and Network, China [2]
Email: {Chunyu.Lin, Jan.DeCock, Peter.Lambert, Rik.VandeWalleg}@ugent.be, yzhao@bjtu.edu.cn

***Abstract*—Lacking of enough 3D contents has become the bottleneck of 3D technology development. 2D-to-3D conversion can not only solve this problem, but also be compatible with the 2D video format. However, most of the real-time 2D-to-3D conversion methods are limited to some special images/videos. This article proposes a scheme that can be applied to all kinds of images/videos. Firstly, the motion vectors are obtained by using bi-directional motion estimation. Combining the motion information and the RGB values of the image, a new image segmentation technology is proposed. With each segmented region, the main vertical objects are detected. Finally, the main vertical objects are assigned the same disparity value based on its bottom vertical position. The disparities of other pixels are determined on its vertical position, which can be seen as a safety methods that uses 3D cues implicitly. The experimental results show its stronger 3D effect.**

***Index Terms*—2D to 3D, disparity assignment, depth segmentation**

I. Introduction

Compared with traditional 2D video, 3D video technology brings much more real and immersive visual experience. The 3DTV is taken as a new revolution for the TV market. Moreover, the development of 3D technology also affects the conventional 2D film, video game, video conference and remote video classroom. The famous 3D film *Avatar* further pushes forward the developing of 3D technology. Currently, a lot of electronic devices have its 3D version, such as 3DTV, 3D computer, 3D mobile phone. However, the development of 3D technology not only depends on the 3D display technology. 3D content is also very important, especially at the pre-3D era. It can be imagined, customers will not upgrade their 2DTV into 3DTV if there are not enough 3D programs to watch. In the meantime, the complicated and time-consuming production process of 3D program results in the lack of 3D content, which becomes the bottleneck of the development of 3D technology.

To solve the problem of lacking 3D content, 2D-to-3D conversion is an effective solution. An efficient 2D-to-3D conversion system can generate the 3D content from existing 2D videos. With such technology, a 3DTV can select to convert the 2D program to 3D form. Hence, it is compatible with 2D video format. Another advantage is that it is impossible to record a lot of old 2D contents into 3D form, while 2D-to-3D conversion can make it simple.

According to the theory of stereo vision, 3D effect is generated because the two images captured by the left and right eyes are different. Our brain integrates these two images into a single 3D image, allowing us to perceive depth information. The distance between the corresponding pixels in the two images is called disparity. With some simple calculation, depth can be easily obtained from the know disparity, or vice versa. Hence, the task of 2D-to-3D conversion is generating a disparity map or depth map. With the disparity map or depth map, together with the original 2D image, two views for our two eyes can be obtained, thus into the 3D format.

There are mainly three classes of 2D-to-3D conversion approaches, which are manual, semi-automatic and automatic ways. Among these three classes, manual approach can get the best results, however, it is time-consuming and too much human labor involved. For example, 3D version of *Titanic* film takes 60 weeks with 300 human workers. Semi-automatic approach can save a lot of human intervention, however, it cannot be processed on real-time. Hence, the automatic scheme, with its feature of real-time and no human intervention, gets a lot of attentions from the research and industry.

Most of the 2D-to-3D conversion algorithm is trying to detect the depth or disparity by exploiting the 3D cues existed in the 2D image/video. In [1] and [2], depth from Defocus(DFM) is employed, in which the extent of blur in the image is used to detect the depth. Generally, the object farther away from the focus of the camera appears with more blur. This kind of scheme requires known parameters of the camera, otherwise more pictures should be captured with different focus setting. Geometric cue includes the familiar size/height of objects in the image, vanishing lines and vanishing points etc., while the vanishing lines and vanishing points are widely used [3]. Vanishing lines refers to the fact that parallel lines, such as railroad tracks, appear to converge with the distance, eventually reaches a vanishing point at the horizon. Based on vanishing lines slope and origin of the vanishing lines, the depth can be estimated. Motion cue refers to the relative motion between the viewing camera and the observed scene, which can be seen as a form of “disparity over time”. At a certain extent, the motion vectors can be directly interpreted as disparity [4], [5]. There are other cues, such as atmosphere cue, texture gradient etc. that can be used. In [6], there is an

978-1-4673-2197-6/12/$31.00 ©2012 IEEE

overview about the these cues.

All the above methods require certain cue existed in the image/video to get the depth. It is easy for our brain to integrate all the cues together and generate a vivid 3D form. However, it is not a simple job for a computer. Hence, we propose a scheme based on main object detecting with tilting method, in which the tilting will generate a safety 3D effect for all kinds of sequence, while main object detecting can correct some wrong disparity values for the main interested objects and enhance 3D effect.

II. Proposed scheme

In this section, the proposed tilting method is introduced firstly. After that, the main vertical object detecting method is introduced to improve the performance.

A. Tilting algorithm

Most of the images/videos are captured with a normal angel. For these images/videos, there is one common feature that the above part of the image/video is the sky(for outdoor) or roof(for indoor), while the bottom part is the ground(for outdoor) or floor(for indoor). In conclusion, from top to bottom of an image/video, it reflects the distance from far to close. Based on such prior knowledge, a safety tilting algorithm is proposed, in which disparity assignment will only consider the vertical position of each pixel. The algorithm is shown in **Procedure 1**.

Procedure 1: Tilting algorithm

```
Input: $img[r, c]$, $degree$;
Output: $oImg[r, c]$;
for $i = 1 : r$
  $t = sin(degree)$;
  $oImg(i$, round($t(r - i)$):$c$)=$img(i$, 1:$c$-round($t(r - i)$));
  $oImg(i$, 1:round($t(r - i)$))=0;
end
```

Where img is the original 2D input image, r and c are the rows and columns of the image. After tilting, the output image is saved in $oImg$. The tilting value for each pixel depends on the vertical position, which are represented by a parameter $degree$. In detail, the sine value of each pixel's vertical position will be taken as the disparity value. The larger the $degree$ is, the larger the disparity will be. Hence, by tuning the $degree$, different extent of depth sense can be obtained. In our case, $degree$ is fixed as 2° throughout of the paper. From the algorithm, it can be seen that the top of the image will have larger disparity value and the lower part of the image will have smaller disparity value. As for the bottom of the image, there will be no disparity at all. In our configuration, the zero plane will be set as on the screen, which means the closest point will be on screen, while the other points are inside the screen. With the original image and the tilted one as a stereo input format, the whole 3D effect is that the image will shift inside the screen from bottom to top. Notice, some pixel at the left border are set as zero when it tilts to other places, which can be seen from the experimental results. Normally, these pixels should be compensated with it neighbor pixels or with a more complicated padding algorithm. For simplicity, these pixels will be set zero. Since, the tilted image will only be shown for one eye and the $degree$ is not large, it is less noticeable when watching it in 3D format.

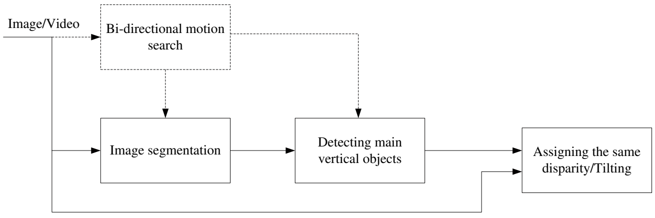

Image/Video
Bi-directional motion search
Image segmentation
Detecting main vertical objects
Assigning the same disparity/Tilting

Fig. 1. The diagram of the proposed scheme.

Notice that not all of the images/videos, or all the pixels of the image can meet this prior knowledge. However, our brain can correct the wrong disparity assignment without any uncomfortable experience. Hence our tilting algorithm actually uses all the 3D cues in the 2D image with an implicit way. The algorithm is very simple and can be implemented real-time. Most important of all, it can be applied to all kinds of images/videos. However, the big problem for such a safety method is that 3D effect is not very noticeable.

The problem of the proposed tilting algorithm is that it just assigns the disparity value based on the vertical position without differentiating the semantic meaning of each pixel. However, it is not easy to segment the image into semantic regions. Moreover, the segmentation will be error prone. In fact, the appropriate disparity/depth value for even one or two objects can increase the 3D performance greatly, especially if the object is the interest region.

Hence, as a compromise, we propose to detect the main vertical objects from a segmented image and take them as a whole to assign the disparity. Furthermore, the motion information is taken into consideration to improve the performance of image segmentation. The whole algorithm is shown in Fig. 1 that is composed of bi-directional motion estimation, image segmentation, main vertical object detecting and tilting as the main steps.

B. Image segmentation with motion information

The approaches based on image segmentation are widely used in stereo matching [7], which assume the same region have the same depth value. In our proposed scheme, the image segmentation is an important step. Hence, the classical algorithm based on graph cut is employed here [8]. In the segmentation algorithm, each pixel is taken as the vertex and the lines between two neighboring pixels are regarded as edges. The weight of each line depends on the difference of the two connected pixels' values. Hence the segmented regions will be composed of vertices and lines based on certain rules. The rule is that the difference between pixels should be small in the same region, while it should be large for the connected pixels in two different regions. Even though the image segmentation algorithm based on graph cut is very efficient [8], the error cannot be avoid, which is a typical problem for image segmentation. Generally, the segmentation rule is just based on the color value, such as [8]. In practice, two regions with the same colors could have different depth. The algorithm, however, cannot differentiate this case. To improve the segmentation results, the motion information in the video is exploited here.

In our scheme, adaptive rood pattern search(ARPS)[9] is employed to get the motion information. Since the motion information in an image is correlated, ARPS employs this knowledge. If the neighbor macroblocks of the current macroblock move at the same direction, then the current macroblock will also moves at the same direction with high probability. With low complexity, ARPS can achieve very good performance, which is the reason to be employed. Here 4×4 block is selected as the motion search unit to get more texture information. Notice that the motion vector obtained by minimizing the mean absolute difference(MAD) may not be the true motion information, especially for the repetitive texture in the image. Hence, we consider the motion vector and MAD together as following

$$cost = MAD + \lambda * mv \tag{1}$$

Where MAD is the mean absolute difference between the current block and the reference block. mv is the mean squared of two motion vectors at horizontal and vertical direction. The parameter $\lambda = e^{-\alpha * mv}$, where α is an empirical value selected as 0.9 in the paper. Hence, mv will have certain weight when mv is small. When mv is larger, MAD will take almost all the weight. Through (1), the wrong motion information for the repetitive structure can be reduced at certain extent.

In a video sequence, most of the objects in the image will move at the same direction. When objects move in one frame, there will be some occlusion appeared in the neighbor frames, or some disocclusion appears. Hence, the single directional motion estimation cannot find the correct motion vector for some blocks. To solve this problem, a bi-directional motion estimation method is used here, just as the B frame in the video coding standard. If the cost of using forward motion vector and backward motion vector have the similar values, the average motion vector will be used. Otherwise, select the vector with the minimum cost. In addition, the occlusion problem can be solved at certain extent. For example, if the cost of using forward motion vector is much smaller than that of backward version, it means there could be occlusion between the current frame f_n and the previous frame f_{n-1}. Hence, the occluded pixels can be compensated by using the corresponding pixels from the next frame f_{n+1}.

Finally, the obtained motion vector will be combined with the segmentation algorithm. The weight of each line is changed as

$$\begin{aligned} w(p,q) = 0.7((r(p) - r(q))^2 + (g(p) - g(q))^2 \\ +(b(p) - b(q))^2) + 0.3(mv(p) - mv(q))^2 \end{aligned} \tag{2}$$

Where $w(p,q)$ denotes the weight connected pixel p and q. r, g and b denotes the luminance value of each color channel. Due to the combination with motion information, the segmented region will be more accurate. When only one image can be used or no motion existed in the video, mv will be set as zero, which degenerates into the typical graph cut segmentation algorithm.

C. Main object detecting algorithm for improving the 3D performance

Even with the proposed segmentation algorithm, not all of the regions can be differentiated correctly. However, we will just focus on main segmented regions, denoted as main vertical object detecting. The main means that the object covers a large area in the whole image. Normally, the object closer to the camera has a large size. In addition, this kind of object is the interested object with high probability because it is put into the focus by intention generally. The vertical means that the object will cover certain larger size on vertical axis. Such a vertical object will be assigned the same disparity/depth value so that the wrong value with the proposed tilting algorithm can be corrected. With the appropriate disparity/depth value for the main interested objects, 3D performance can be improved a lot.

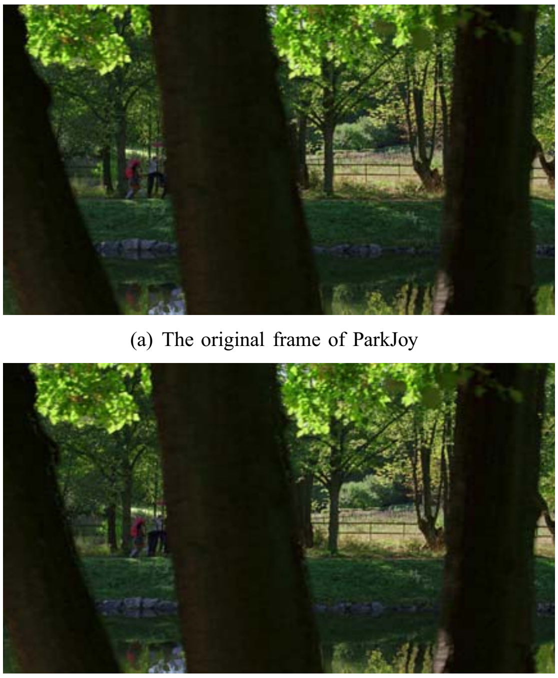
In practice, the main vertical object is defined as an area larger than $[0.1 * cc, 0.25 * r]$, which is an empirical value. The object smaller than this size is not the main interested object with high probability. Or the tilting algorithm can also generate certain 3D effect. In addition, if there is motion, the object closer to the camera should have larger motion vector. This factor is also considered to exclude some wrong detected objects. To make the detecting fast and accurate, the parameter for over segmentation is applied so that a lot of small regions will not be generated and considered.

With such constraint, the main vertical object can be detected more accurate. For the detected object, the same disparity value will be assigned based on its bottom vertical position. For the other objects, tilting algorithm will be used. The whole algorithm can provide safety results for all kinds of images/videos. As for the images/videos that contain main vertical objects, the wrong disparity/depth can be corrected and enhanced 3D effect can be obtained further.

III. Experimental results

In this section, the experimental results are shown. The HD sequence *ParkJoy* and SIF sequence *Garden* are used for the testing. For comparison, the original images are also shown. To make the difference more noticeable, we do not do any processing here. For example, the black region on the left border can be noticed clearly in Fig.3. In practice, the occlusion after the disparity/depth assignment should be compensated. In the Fig.2 and Fig.3, the trees are the main detected objects with the same assigned disparity value as a whole, while the disparity of other regions is obtained with our tilting algorithm. Hence, the detected tree will have more 3D effect, which enhances the performance of the whole scene. More results can be found at our website[1], where you can

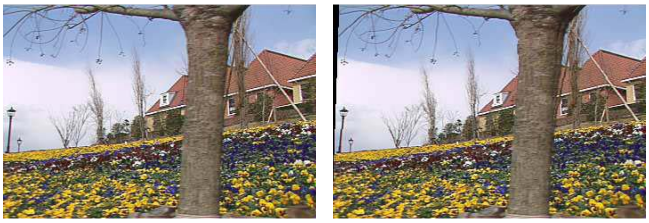
[1] http://multimedialab.elis.ugent.be/users/chlin/Depth_map_results



(a) The original frame of ParkJoy

(b) The obtained frame of ParkJoy

Fig. 2. Stereo format with left and right image for ParkJoy



(a) The original frame of Garden (b) The obtained frame of Garden

Fig. 3. Stereo format with left and right image for Garden

select the original image and the generated image as left and right view to be displayed on a 3DTV. Fig. 4 gives a depth map calculated from the obtained disparity map in order to show an intuitive result. Notice there is not any processing yet, such as occlusion removing. This result shows that main vertical object will be given an more accurate depth. Even the depth value for other pixels may not be accurate, however, it can be corrected with our brain. It is not easy to provide a more objective results for 3D effect, so the best way is downloading the material from our website and watching them with a 3D displaying device.

IV. Conclusion

In this paper, an automatic 2D-to-3D conversion algorithm is proposed, which is composed of safety tilting, segmentation with bi-directional motion estimation and main vertical object detecting. For the detected main vertical object, the same disparity value, calculated with its bottom vertical position, is assigned for the object as a whole. For the other regions, tilting algorithm is used based on each pixel's vertical position. The tilting algorithm is a safety method that can be applied to all kinds of images, with less noticeable 3D effect. Even not all the pixels can be given a correct disparity value, our brain can correct the error using the implicit 3D cues in the image. Combined with appropriate disparity assignment for the main vertical object, 3D effect is improved a lot.

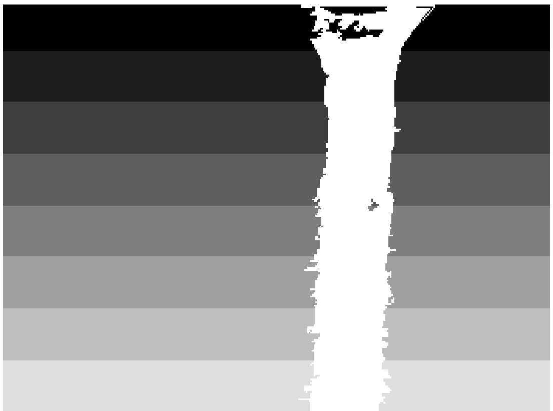


Fig. 4. The converted depth for Garden

Acknowledgement

This research described in this paper was carried out as part of the IBBT/IWT 3DTV2.0 project and was funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

References

[1] S. W. Hasinoff and K. N. Kutulakos, "Confocal stereo," *Int. J. Comput. Vision*, vol. 81, no. 1, pp. 82–104, Jan 2009.

[2] S. Zhuo and T. Sim, "On the recovery of depth from a single defocused image," in *Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns*, ser. CAIP '09, 2009, pp. 889–897.

[3] S. Battiato, S. Curti, M. L. Cascia, M. Tortora, and E. Scordato, "Depth map generation by image classification," B. D. Corner, P. Li, and R. P. Pargas, Eds., vol. 5302, no. 1. SPIE, 2004, pp. 95–104.

[4] M. T. Pourazad, P. Nasiopoulos, and R. K. Ward, "Generating the depth map from the motion information of H.264-encoded 2D video sequence," *J. Image Video Process.*, vol. 2010, pp. 4:1–4:13, January 2010.

[5] D. Kim, D. Min, and K. Sohn, "Stereoscopic video generation method using motion analysis," in *Proc. 3DTV Conf*, 2007, pp. 1–4.

[6] L. Zhang, C. Vazquez, and S. Knorr, "3D-TV content creation: Automatic 2D-to-3D video conversion," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 372 –383, June 2011.

[7] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *18th International Conference on Pattern Recognition, ICPR 2006.*, vol. 3, 2006, pp. 15–18.

[8] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, pp. 167–181, September 2004.

[9] Y. Nie and K.-K. Ma, "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Transactions on Image Processing*, vol. 11, no. 12, pp. 1442 – 1449, Dec 2002.