Stereoscopic Learning for Disparity Estimation

Zhebin Zhang*,† Yizhou Wang†,◇ Tingting Jiang†,◇ Wen Gao†,◇, *Fellow, IEEE*
*Key Lab. of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
*Graduate School, Chinese Academy of Sciences, Beijing, 100039, China
†National Engineering Lab. for Video Technology,
◇Key Lab. of Machine Perception (MoE), School of EECS, Peking University, Beijing, 100871, China
zbzhang@jdl.ac.cn, {yizhou.wang, ttjiang, wgao}@pku.edu.cn

***Abstract*—In this paper, we propose a learning based approach to estimating pixel disparities from the motion information extracted out of input monoscopic video sequences. We represent each video frame with superpixels, and extract the motion features from the superpixels and the frame boundary. These motion features account for the motion pattern of the superpixel as well as camera motion. In the learning phase, given a pair of stereoscopic video sequences, we employ a state-of-the-art stereo matching method to compute the disparity map of each frame as ground truth. Then a multi-label SVM is trained from the estimated disparities and the corresponding motion features. In the testing phase, we use the learned SVM to predict the disparity for each superpixel in a monoscopic video sequence. Experiment results show that the proposed method achieves low error rate in disparity estimation.**

I. Introduction

The film *AVATAR* starts a new era of 3D entertainment industry. However, the production of 3D videos becomes a bottleneck of the 3D movie and 3DTV industry. Although new videos can be produced directly by 3D cameras, the market still has huge demand for converting old monoscopic videos into stereoscopic ones. The major challenge of 2D-to-3D video conversion lies in the disparity estimation of monocular videos [1]. There have been a number of approaches about 2D-to-3D video conversion. These methods can be classified into two categories. One class is fully automatic conversion [9] [11] [10] [12]. These methods use motion to predict the depth/disparity of pixels. Some commercial softwares, such as DDD's TriDef 3D player and Samsung's 3DTV, can generate stereoscopic views from monocular videos in realtime. However, these products are not able to robustly estimate pixel disparities, neither consider the inward and outward visual perception effects (see Figure 1). The other category of methods leverage user interactions during stereoscopic conversion [6]. At key frames, they draw scribbles to initialize the depth values of the objects in the scene, and then automatically propagate the depth information to the unlabeled frames. Although this type of methods achieve better performance than those fully automatic ones, it is still very labor intensive to label large amount of videos. Moreover, even for a single person, his/her disparity labeling can be inconsistent or unreliable.

Estimating pixel disparities from monocular videos is the key step in 2D-to-3D video conversion. In this paper, we are interested in the automatic stereoscopic conversion methods, especially using motion cue. Under the condition of constrained camera motion and assuming that scenes are static, there exist two ways to estimate disparity maps, (i) using Structure from Motion(SFM) [7] [11] [10] [12], and (ii) leveraging motion parallax [9]. However, in real scenario such as movies and TV programs, the constrained camera motion condition and static scene assumption are often violated, which

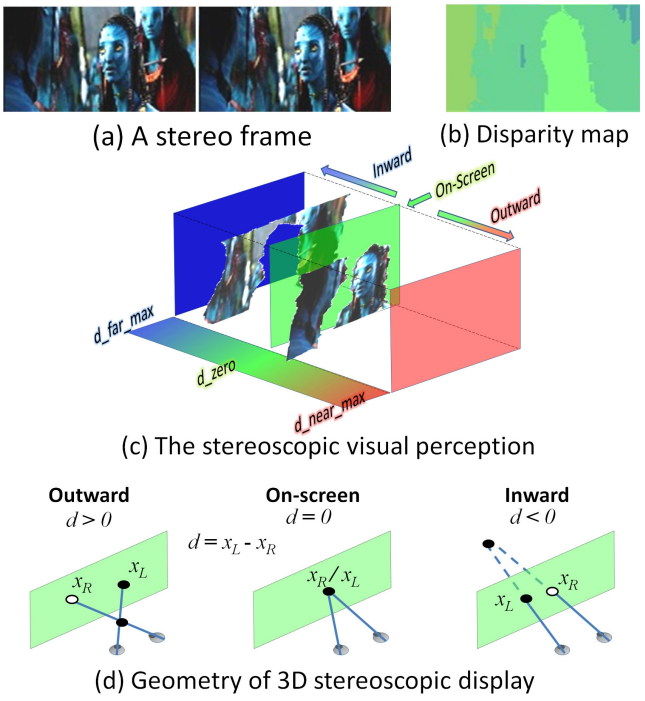

Fig. 1. **The principle of 3D stereoscopic display (best viewed in color)**. When we see a 3DTV, there exists slight difference between a pair of images captured by the left eye and the right eye **(a)**, namely disparity, because of which the stereoscopic visual perception arises. The disparities of pixels may vary due to their different depths in the 3D scene, as shown in the disparity map **(b)**. Disparity values have signs (**(d)**), which decide three different stereoscopic visual perceptions: **outward perception (positive), on-screen perception (zero) and inward perception (negative)**. The absolute value of disparity of a pixel decides how far the pixel is perceived from the screen. As shown in **(b)**, we visualize the disparity map and use green color to indicate the zero disparities, blue for the negative disparities and red for the positive ones. The deeper the color, the farther a pixel away from the screen **(c)**. In this paper, we propose a method using a multi-label classifier to predict pixel disparities.

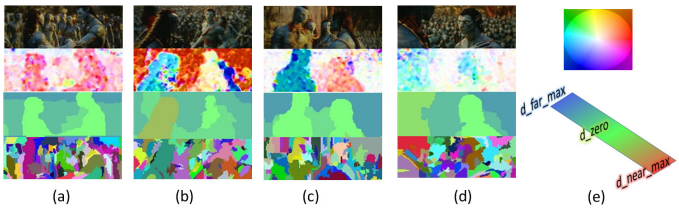

Fig. 2. **Motion and Disparity.** We visualized the motion(optical flow) and disparity map according to a value-orientation-to-color transform (top of **(e)**) and a value-to-color transform (bottom of **(e)**). The relation between disparity and motion is not absolutely proportional. **(a)**Similar motion and similar disparity. **(b)**Different motion and different disparity. **(c)** Different motion but similar disparity. **(d)** Similar motion but different disparity

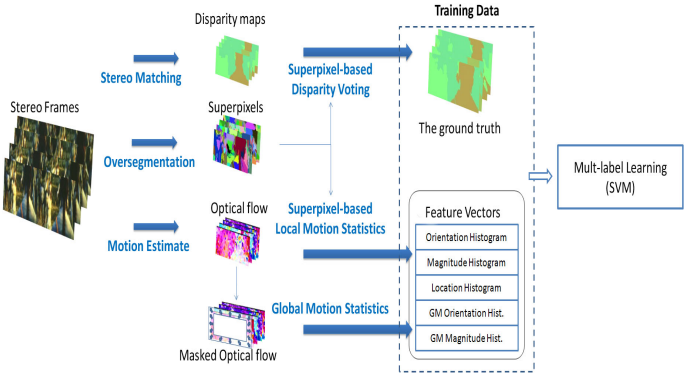

Fig. 3. **The flowchart of the stereoscopic learning.**

leads to the failure of applying the two methods in disparity estimation. For example, as shown in Figure 2 (c) and (d), even with the same motion, the disparity can be different; and similar disparities do not imply similar motions.

In this paper, we propose a learning based approach, which automatically estimates pixel disparities of monoscopic video sequences. According to our experience, we identified two key elements in the 2D-to-3D video conversion: (i) separating pixels into layers of different disparities, and (ii) arranging inward/outward perception effects of the disparity layers (see Figure 1). The absolute values of disparities can be estimated from different types of cues such as motion [9] [10] [11] [12], geometry [4] and appearance [6] [8] [13]. The inward/outward perception relates to cinematography artistic effect as well as comfort of viewing. However, this issue is not considered in the existing 2D-to-3D video conversion systems. However, the recently released stereoscopic movies and TV programs provide us good examples to learn the principles of how to simplify disparity maps as a sufficient number of layers and how to arrange the inward/outward visual effect. In this paper, we simultaneously consider both issues during the 2D-to-3D video conversion.

In the proposed method, we consider both the global motion of a scene(e.g. due to the camera motion) and the local motion of regions or objects in order to estimate disparity maps. We represent each video frame with superpixels, and extract the motion features from the superpixels and the frame boundary. We observe that the content in a frame boundary usually corresponds to scene background. As a result, the motion feature extracted from this region usually tells the camera motion. The global motion or camera motion provides a context, which can be helpful in determing absolute motion (and motion parallax) of the foreground objects and consequently helps to determine their disparities.

In the learning phase, given a pair of stereoscopic video sequences, we employ a state-of-the-art stereo matching method to compute the disparity map of each frame as ground truth. Note that we do not perform parallel view rectification before the stereo matching, hence, the disparity has signed value. The disparity values are quantized into several levels and treated as stereoscopic labels. Then a multi-label SVM is trained from the estimated disparities and the corresponding motion features. It is expected that the trained model can capture the correlation between the motion and the **signed** disparity. In the testing phase, we use the learned SVM to predict the disparity for each superpixel in a monoscopic video sequence.

The contribution of the proposed learning based method is two-fold: (i) it is able to estimate disparities without the constraint of camera motion scenario. (ii) besides the disparity values, the proposed method also arranges the inward/outward perception effects.

The rest of this paper is organized as following. In Section II, we give an overview of the proposed method, followed by a detailed description of the method in Section III. We show the experiment settings and results in Section IV. Section V concludes the paper.

II. Overview of the Approach

We adopt a supervised learning scheme to study the correlation between the disparity and the motion features introduced in Section I. Figure 3 shows the flowchart of the proposed stereoscopic learning method. In the training phase, we first oversegment the reference image of the stereo image pair into superpixels and use them as atomic elements of our image representation. Based on this representation, we extract the motion feature by calculating the statistics of optical flows in each superpixel and in the boundary regions of the image(as shown in Figure 3 and Figure 4) in terms of optical flow histograms. Meanwhile, to obtain the ground truth, we adopt a state-of-the-art stereo matching method [14] to compute pixel-wise disparities on the image pairs. For each superpixel, its stereoscopic label is assigned as the most voted disparity value in the superpixel. In this work, we do not rectify stereo pairs into parallel views in order to obtain signed disparities. This is reasonable since all the videos used in this paper are movies and TV programs shot professionally, and we can safely assume that the corresponding pixels are on the same horizontal line. Hence, the proposed method can predict the pixel disparities and render inward and outward visual perception effects simultaneously.

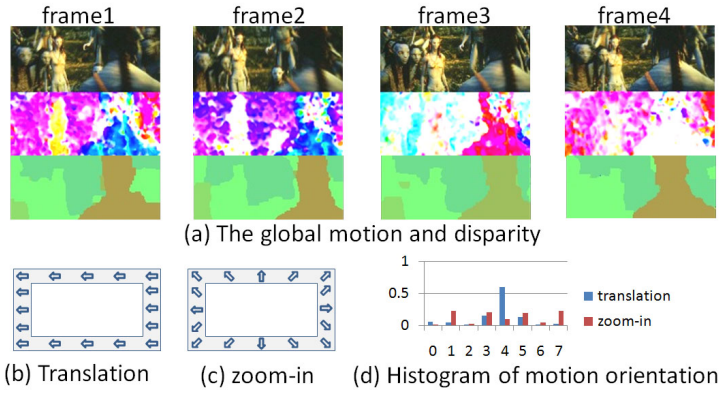

Fig. 4. **The motion feature and its motivation. (a)** Local motion changes with the global motion among four consecutive frames although the disparity maps are similar. **(d)** Histogram of motion orientation for the frame's inner boundary regions for different global motion patterns, such as translation **(b)** or zoom-in **(c)**.

III. Stereoscopic Learning

A. Stereoscopic display and stereo perception

In order to have stereoscopic perception, the display device needs to send each of the two views to the corresponding eye. Active 3D display and passive 3D display methods have been invented using shutters, polarization or anaglyph glasses [1].

As shown in Figure 1, it is the disparity that generates 3D perception. Let $d = x_L - x_R$ denote the disparity for a corresponding pixel pair on the same scan line, where x_L and x_R are the horizontal coordinates of the pixel pair on the left image and right image respectively. The signed disparity d creates different stereo perception effects: (1) **Negative disparity** ($d < 0$) generates the inward screen perception, (2) **Zero disparity**($d = 0$) generates the on-screen perception, and (3) **Positive disparity**($d > 0$) generates the outward screen perception. The absolute value of d decides how far the perceived depth value of a pixel is from the screen. Based on this observation, in this paper, we directly use the disparity to represent the stereoscopic information instead of the depth value.

B. Motion features

In section I, we discuss the relation between the motion and the disparity. From Figure 2, we can see that the correlation between the optical flow (the motion) and disparities is not an exact proportional relation as used in [9], especially when there exists the global motion of the scene, such as pan, zooming in and out. As shown in the Figure 4 (a), the foreground object motion can be greatly impacted by the global motion. Hence, when using the motion cue to estimate disparities, we need to consider local and global motion together. In this paper, we concatenate global motion and local motion as a joint feature. Figure 4 illustrates the idea. For the global motion, we compute histograms of optical flows on the boundary of the frame, including the orientation histogram and magnitude histogram. A computed histogram of optical flow orientation is shown in Figure 4 (d). The width of the boundary region is one eighth of the frame width. It is expected that the statistics of motion in the boundary region can reveal the underlying global motion pattern. In this way, we avoid the camera tracking used in SfM [7]. In addition, to obtain a smooth disparity of a superpixel in temporal domain, we include overlap components in the motion features of the corresponding superpixels in two adjacent frames. To do so, both the forward and backward optical flow histograms are concatenated into the feature vector.

As shown in the Figure 3, the feature is composed of the histograms of quantized orientations and magnitudes of optical flows of the local motions and global motions. Figure 2 (a) and (b) shows that the same motion pattern in a frame does not always result in the same disparity. So we have to resort to other cues in order to estimate more accurate disparities. Inspired by [8], we also concatenate the location of the superpixel into its feature vector. The superpixel's centroid coordinates are normalized by the horizontal and vertical dimension of the image.

C. Multi-label classifier learning

We cast the disparity prediction as a multi-label classification problem. Each quantized disparity value is taken as a stereoscopic label of the superpixel as discussed in III-A. A supervised multi-class learning scheme is used to establish the relation between the feature and the stereoscopic labels.

1) Assigning disparity labels: To obtain the ground truth of stereoscopic labels for both training and testing, we first over-segment the left image of the stereoscopic pair into superpixels [5], and compute its disparity map using the state-of-the-art stereo matching method [14]. In the stereo matching phase, in order to enhance the stereoscopic perception effects of the video, we allow the both positive and negative values in the same disparity map. To assign a stereoscopic label for a superpixel, we use the mostly supported disparity value d_{r_s} in the superpixel r_s as its label.

$$d_{r_s} = \arg \max_{d_i \in D_{r_s}} \#(d_i) \tag{1}$$

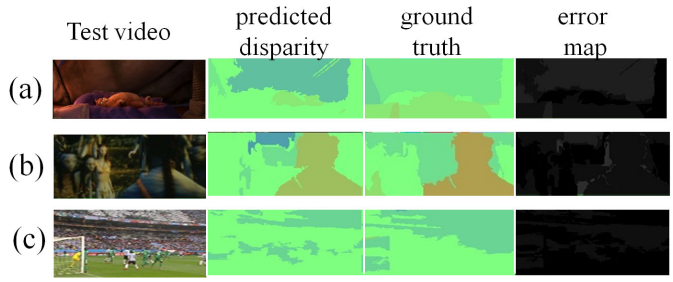

Fig. 5. Example results on 3 different test videos. (a) Shrek. (b) Avatar. (c) FIFA World Cup 2010.

where d_i is a quantized disparity value in the superpixel, $\#(d_i)$ is the number of pixels whose disparities equal to d_i, and D_{r_s} is the set of disparity values in appeared r_s.

2) Training: A multi-class SVM [2] is trained to predict pixel disparities from the motion features. In our implementation, we use LLIBSVM [3] and the one-vs-all strategy to train the multi-label classifier.

IV. Experiments and Results

A. Expriments

1) Stereoscopic video dataset: We collect a set of stereoscopic videos from the company Tru3D's website (http://www.tru3d.com) and some other internet resources. The collection includes the movie trailers of "Shrek" and "Avatar", and also includes video clips of "FIFA World Cup 2010".

2) Training Data: We use the method in [5] to do the over-segmentation, where we set the parameters as $\sigma = 0.5$, $k = 500$, $min = 20$. Even if the accuracy of disparities of the movie data is not as good as that of the Middlebury stereo matching evaluation benchmark [15], it still can provide us enough depth/disparity information for stereoscopic perception. In the implementation, we use the Graph Cuts algorithm [14] for stereo matching. The minimal disparity value is set as $-D_{Range}$ and the maximal disparity value is set as to D_{Range}, where $D_{Range} = W_{frame}/20$ and W_{frame} is the video frame width. Figure 2 and Figure 4 (a) show some examples of the training data and the ground truth label.

B. Results

We use different sets of data for training and testing. Figure 5 shows some examples of the estimated disparity maps on "Avatar," "Shrek" and "World Cup 2010." We use red, green and blue colors to indicate the outward perception, on-screen perception and inward perception respectively (as introduced in the Figure 1). The results are compared to the ground truth. The disparity estimation error map for each example (the gray image) are shown. To evaluate the sensitivity of disparity quantization to the estimation result, we quantize the disparity value range into 7, 9 and 11 levels. Table I shows the performance on the three datasets with different disparity level numbers. From the results, we can see that the disparity estimation accuracy is reasonably stable across different number of quantizations.

TABLE I
Disparity estimation error with respect to different disparity level numbers

Disparity level number	FIFA	Shrek	Avatar
7	15.7%	14.3%	17.9%
9	15.0%	14.5%	18.8%
11	15.5%	15.0%	18.1%

V. Conclusion

In this paper, we utilize machine learning techniques to capture the relation between the disparity and the motion by which we can predict the disparity maps for the monoscopic video. The experimental results show that the proposed method is able to estimate disparities of general scenario without constraints of camera motion. Besides the disparity values, the proposed method also arranges the inward/outward perception effects.

VI. Acknowledgement

We'd like to thank for the support from research grants NSFC-60872077, 973-2009CB320904 and Doctoral Fund of MoE KEJ200900029.

References

[1] 3D4YOU. 3D4YOU project, WP2: Requirements on post-production and formats conversion, 2008. Philips, CAU, KUK, BBC, TMMRDF, HHI.
[2] C. Burges. A tutorial on support vector machines for pattern recognition, 1998. Knowledge Discovery and Data Mining.
[3] C.-C. Chang and Chih-Jen. A library for support vector machines, 2000. http://www.csie.ntu.edu.tw/ cjlin/libsvm/.
[4] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology, 2000. IJCV.
[5] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation, 2004. IJCV.
[6] M. Guttmann, L. Wolf, and D. Cohen-Or. Semi-automatic stereo extraction from video footage, 2009. ICCV.
[7] R. I. Hartley and A. Zisserman. Multiple view geometry in computer vision, 2000. Cambridge University Press.
[8] D. Hoiem, A. A. Efros, and M. Hebert. Closing the loop on scene interpretation, 2008. CVPR.
[9] D. Kim, D. Min, and K. Sohn. A stereoscopic video generation method using stereoscopic display characterization and motion analysis, 2008. IEEE TRANSACTIONS ON BROADCASTING.
[10] S. Knorr and T. Sikora. An image-based rendering IBR approach for realistic stereo view synthesis of tv broadcast based on structure from motion, 2007. ICIP.
[11] K. Moustakas, D. Tzovaras, and M. Strintzis. Stereoscopic video generation based on efficient layered structure and motion estimation from a monoscopic image sequence, 2005. Circuits and Systems for Video Technology.
[12] E. Rotem, K. Wolowelsky, and D. Pelz. Automatic video to stereoscopic video conversion, 2005. SPIE.
[13] A. Saxena, M. Sun, and A. Y. Ng. Make3D: Learning 3-D scene structure from a single still image, 2008. PAMI.
[14] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithm, 2002. IJCV.
[15] D. Scharstein and R. Szeliski. Middlery stereo benchmark, 2005. http://vision.middlebury.edu/stereo/eval/.