Object Segmentation for Z-keying Using Stereo Images

Woontack Woo, Namgyu Kim and Yuichi Iwadate
ATR MIC Labs
Kyoto 619-0288, Japan
Fax: +81 (774) 95-1408, E-mail:{wwoo,ngkim,yiwadate}@mic.atr.co.jp

Abstract:
In this paper, we propose a robust object segmentation framework exploiting multiple cues such as shape, intensity (color) and depth. Though over last few decades, various segmentation schemes have been developed, those schemes based on intensity and motion information have well-known disadvantages. To alleviate those problems we take into account depth information using MRF/GRF framework. The experimental results show the effectiveness of the proposed framework by clearly separating "objects of interest" from real, rather than blue-screen, scenes. The proposed scheme will be a key part for wide scope of applications requiring object-based functionalities as well as z-keying for photo-realistic mixed reality.

Keywords:
object segmentation, stereo images, z-keying

1 Introduction

Over last few decades, there has been growing interest in object segmentation to provide various applications with object-based functionalities such as interactivity and scalability. Nevertheless, coming standards, *e.g.* MPEG-4 and MPEG-7, do not specify how to segment objects, even though those standards highly rely on object-based representation. The difficulties of automatic segmentation come mainly from defining *semantically meaningful areas* because the definition of *meaningful areas* itself is not so clear in many cases [1]. Even in case we have clear definition about the areas of interest, an accurate and reliable segmentation is a challenging and demanding task because those areas are not homogeneous with respect to low-level features such as intensity, color, motion, disparity, etc.

To alleviate those difficulties, several hybrid schemes have been proposed [2–5]. Among those, motion information has been widely accepted as a crucial cue, under assumption that the objects of interest (OOI) can be characterized by a coherent motion. Note however that this scenario only works for the case of that the OOI have motion in the scene. Even in case of moving objects, motion similarity may not work well due to various error sources including occlusions and inaccurate motion estimation. Therefore, additional information is inevitable to detect accurate boundaries of objects.

In this paper, we propose a robust segmentation scheme jointly exploiting edge and disparity information. The proposed scheme consists of three steps, which are (1) edge detection and smoothing, (2) depth estimation and (3) object segmentation. We first estimate intensity edge and smoothen both images to reduce noise effects. We use resulting edge information as an initial disparity edge based on the assumption that pixels tend to have similar disparity, as well as intensity, within a rigid object. Then, given intensity edge and smoothed image pair, we estimate a consistent disparity field with sharp boundary using proposed MRF/GRF framework [6–9]. Finally, given disparity and boundary information, the OOI is separated from the target image, under assumption that the OOI has limited range of depth and the disparity is smoothly changing within the OOI, in general. The resulting OOI containing depth information is ready to be mixed into another image/video using z-keying, which is a key part of the photo-realistic mixed reality system.

This paper is organized as follows. In Section 2 the proposed hybrid segmentation framework is described more in detail. Some experimental results and possible extension of this research are given in Section 3.

2 Hybrid Object Segmentation

2.1 Problem Formulation

Let F_1 and F_2 be, respectively, the segmentation target and reference images in a stereo pair. The image can be represented as a set of pixels, *e.g.* $F_1 = \{f_{ij}^1, 0 \leq i < N_x, 0 \leq j < N_y\}$, where f_{ij} de-

0-7803-5747-7/00/$10.00

notes (i,j)th pixel and the index 1 represent the target image. The number of pixels is $N_x \times N_y$, where N_x and N_y, respectively, represent the height and width of the image. The resulting pixelwise disparity vector field (V) can be represented as $V_1 = \{v_{ij}, 0 \leq i < N_x, 0 \leq j < N_y\}$. Then, $\hat{f}^1_{ij} = f^2_{ij \oplus v_{ij}}$, where $\oplus$ denotes the displacement.

In general, except object boundaries, intensity and disparity fields are generally very smooth, in fact disparity fields are much smoother than images themselves. Thus, as we did in [9], both image and disparity field can be modeled using MRF model with appropriate *a priori* assumptions on its smoothness. Figure 1 shows pre-defined first order neighborhood. We define two types of neighborhood systems, *i.e.* normal and causal, where each is used in pixelwise and hierarchical blockwise operations, respectively.

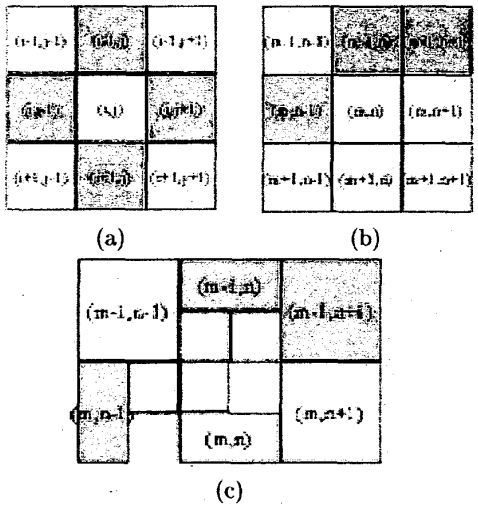


Figure 1: First order neighborhood system. (a) Normal neighborhood (b) Causal neighborhood (c) Hierarchical Causal neighborhood for the up-left block of the (m,n)*th* block. Two types of neighborhood systems, *i.e.* normal and causal, are defined where each is used in pixelwise and hierarchical blockwise operations, respectively. The indexes (i,j) and (m,n) denote pixel and block, respectively.

For a given stereo pair, F_1 and F_2, we formulate the problem as follows. We want to estimate $\hat{X}$, *i.e.* object of interest, O_1, disparity contour, L_{V_1}, disparity, V_1, intensity contour, L_1, and smooth stereo pair, G_1 and G_2, such that $\hat{X}$ maximizes the *a posteriori* probability (MAP), $P(O_1, L_{V_1}, V_1, L_1, G_1, G_2 | F_1, F_2)$. The *a posteriori* probability is decomposed using Bayes theorem and then the MAP estimation is replaced by the energy minimization problem according to the Clifford-Hammersley theorem, the theory of equivalence of MRF and Gibbs random fields [10].

Along with the similar process as we did in [6–9], we define the MAP estimation as follows.

$$\begin{aligned}
\hat{X} &= \arg\max_X P(O_1, V_1, L_{V_1}, G_1, L_1, G_2, L_2 | F_1, F_2) \quad (1) \\
&= \arg\max_X P(O_1 | V_1, L_{V_1}) \cdot P(V_1, L_{V_1} | G_1, L_1, G_2) \\
&\quad \cdot P(G_1, L_1 | F_1) \cdot P(G_2, L_2 | F_2) \\
&= \arg\min_X U(O_1 | V_1, L_{V_1}) + U(V_1, L_{V_1} | G_1, L_1, G_2) \\
&\quad + U(G_1, L_1 | F_1) + U(G_2, L_2 | F_2)
\end{aligned}$$

where U denotes the energy function. In (1), each term is related to object segmentation, disparity estimation and edge detection, respectively. As explained, we first estimate edge and perform edge-preserved smoothing. Then, given edge information and smooth image pair, we estimate disparity field with sharp boundaries using the derived energy equation based on MRF/GRF framework. Finally, we segment the OOI from the target image, given object contours and disparity vector field.

2.2 Edge Detection

Before estimating disparity field, we estimate intensity edge and perform edge-preserved smoothing for each image of the pair, which corresponds to the third and fourth terms in (1). To avoid local minima in edge detection, we provide an initial discontinuity, the intensity difference, which is decided as

$$l_{ij} = \begin{cases} 1, & |f_{ij} - f^{\eta}_{ij}| \geq T_e \\ 0, & o.w. \end{cases} \quad (2)$$

where T_e is a threshold for edge decision. If the intensity difference between its neighborhood pixel exceeds the threshold T_e, then we assume there is high possibility of discontinuity. Given the initial discontinuity, the cost function for edge detection and smoothing process for the target image can be defined as follows.

$$\begin{aligned}
& U(G_1, L_1 | F_1) \quad (3) \\
= \; & U(F_1 | G_1, L_1) + U(G_1 | L_1) + U(L_1) \\
= \; & \sum_{(i,j) \in \Omega} \{(1 - \alpha_s) \| g^1_{ij} - f^1_{ij} \|^2 \\
+ \; & \alpha_s \sum_{\eta_{ij}} (g_{ij} - g^{\eta}_{ij})^2 (1 - l^{\eta}_{ij}) + \beta_s \sum_{c \in C} V_c(l_{ij}, l^{\eta}_{ij}) \}
\end{aligned}$$

where Ω and C represent a rectangular lattice and a pre-specified set of cliques, respectively. In the above equation, f, g and l represent pixelwise intensities and edge, respectively. The parameters α_s and β_s are weighting constants controlling the weight among similarity, smoothness and discontinuity.

In (3), the first term corresponds to the noise process, which occurs when 3D scene is projected onto 2D

image through a camera lens. The second terms corresponds to the *a priori* assumption on the smoothness of the image, S_1, given intensity edges, L_1. The last term represents a cost function for the intensity edge, which controls the discontinuity of intensity levels between the pixel and its neighborhood. The main role of the edge is to prevent intensity from being oversmoothed across object boundaries, *e.g.* in case $l_{ij} = 1$, the smoothness constraint should not be performed across this discontinuity. In addition, the resulting intensity edge is a good initial guess of the disparity edge, though the intensity discontinuities may not correspond to physical boundaries of the OOI [11].

2.3 Disparity Estimation

In most intensity-based disparity estimation methods, fixed shape of pixels or fixed size of block have been used to measure the similarity between stereo pair. Note however that they usually tend to fail to provide good matching results. For example, as the block size becomes larger, the level of estimation error increases, especially when the block includes object boundaries. By reducing the block size, the estimation error can be reduced but the resulting disparity field may not be homogeneous because the estimation is subject to various noise effects.

Various estimation schemes have been proposed to overcome those drawbacks, which include estimation schemes with MRF model [6], overlapped block matching (OBM) with modified MRF model [12]. Nevertheless, conventional block-based approaches only relieve parts of drawbacks. A way to overcome the dilemma between robustness and consistency of intensity-based disparity estimation is to adopt hierarchical block segmentation or adaptive window in disparity estimation [13–15]. The main advantage of the disparity estimation with hierarchical block segmentation is that it can overcome mismatching problem (inconsistency of the disparity field) by considering a large area during the initial disparity field estimation. However both frequently fail to provide an accurate disparity along object boundaries due to occlusion effects.

To yield a robust and consistent disparity field, we adopt two step approach, *i.e.* hierarchical OBM disparity estimation and then pixelwise refinement. Note that to estimate a smooth disparity field we apply an enlarged overlapped window and modified causal neighborhood during hierarchical block matching. A basic procedure of disparity estimation based on hierarchical block segmentation is as follows. First, the disparity is estimated at the coarsest level, *e.g.* block size of 16×16, using full search block matching. The basic idea of block segmentation is that the block is segmented into smaller blocks, *e.g.*, 16×8 or 8×8, only if the block yields higher compensation error level than the pre-fixed threshold value. However, we ignore the segmentation and keep original disparity, if the reduction of error level is small enough between a block and its consecutive subblocks. Segmentation is repeated until the error of the block is smaller than the summed error of segmented subblocks or the block size is reached the preselected size.

Then, given an initial disparity field, edge information and smooth observations of a stereo pair, (G_1, G_2), we perform pixelwise disparity estimation based on coupled MRF/GRF model with respect to the predefined first order neighborhood system. The solution of second term in (1), corresponding to the cost function of pixelwise disparity estimation, will result in smooth disparity field with sharp boundary. The corresponding cost function $U(V_1, L_{V_1}|G_1, G_2, L_1)$ is defined as follows.

$$\begin{aligned} & U(V_1, L_{V_1}|G_1, G_2, L_1) \\ = \; & U(G_1|G_2, V_1) + U(V_1|L_{V_1}) + U(L_{V_1}|L_1) \\ = \; & \sum_{(i,j)\in\Omega} \{(1-\alpha)\|g^1_{ij} - g^2_{ij\oplus v_{ij}}\|^2 \\ + \; & \alpha \sum_{\eta_{ij}} (v_{ij} - v^\eta_{ij})^2 (1 - l^\eta_{v,ij}) \\ + \; & \beta \sum_{c\in C} V_c(l_{v,ij}, l^\eta_{v,ij}, l_{ij})\} \end{aligned} \tag{4}$$

where Ω and C represent a rectangular lattice and a pre-specified set of cliques, respectively. In the above equation, g, v and l represent pixelwise intensity, disparity, and contour, respectively. The disparity edge l_v controls the discontinuity between the disparity, v, and its neighborhood, v^η. The parameters α and β are weighting constants controlling the weight among similarity, smoothness and discontinuity. The subscript $\oplus$ represents a movement along the disparity.

In general, the intensity levels in a stereo pair may not be the same, even if the images are captured at the same time and at the same place. In (4), the first term corresponds to the constraint on the similarity of intensity between corresponding stereo images along the disparity. The second terms in (4) corresponds to the *a priori* assumption on the smoothness of the disparity field given disparity edges, L_{V_1}. We assume that the real disparity field is smooth except for the object boundaries that are related to the depth discontinuities. The last term in (4) represents a cost function for the disparity edge, which controls the discontinuity between the disparity and its neighborhood. The main role of the disparity edge is to pre-

vent disparity from being oversmoothed across object boundaries. Thus, smoothness constraint should not be performed across this discontinuity.

For the disparity estimation we only use the first and the second term in (4) because only two terms are direct function of disparity. We estimate disparity by tradeoffs between similarity of intensities and smoothness of the disparity field. Similarly, the disparity edge is decided using the second and the third terms in (4). In this case, the role of the second term is a kind of dynamic thresholding for the decision of an edge according to the state of the neighboring edges. The obtained dense disparity field using (4) is smooth yet has sharp boundary.

2.4 Object Segmentation

At first, we capture background scene without the OOI and then keep it to generate initial OOI segmentation map when the OOI, *e.g.* participants, appear in the target image.

At the final stage, given disparity and edge information, we segment object from the target image using the first term in (1). In general, disparity field contains crucial information about depth structure of the scene, while edge provides shape information. We assume that the pixels with smooth disparity variation belong to the same object. In addition, we assume the OOI has limited range of depth and the depth within the OOI is changing smoothly. In this scenario, a coarse OOI segmentation map can be refined by properly combining disparity and edges.

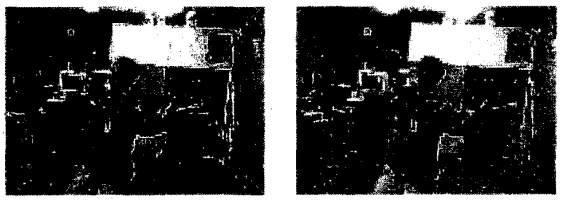
3 Experimental Results

The experimental results illustrate the effectiveness of the proposed hybrid object segmentation scheme using stereo images. At the first, edge is estimated and then edge-preserving low pass filtering is applied to both images, which reduces noise effect and thus helps further processing. Then, given a pair of smooth images and edge information, we estimate smooth disparity field while keeping sharp boundary. Finally, we refine the segmentation map by properly combining several cues such as intensity, intensity edge, disparity and disparity edge.

Figure 2 (a) and (b) show a pair of stereo images used in the experiment, where the size is 240×320.

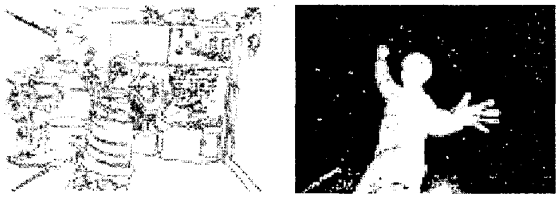
Figure 3 (a) and (b) show resulting binary edge image and initial segmentation map.

We assume the stereo pairs satisfy epipolar geometry and thus restrict search window to horizontal direction, *e.g.* $(0, 32)$ and then estimate the disparity with a minimum energy by full search within a search window [16]. In this experiments we use intensity edge as disparity edge to speed up the processing time. The initial block size is 8×8 and the block with high level of estimation error is segmented into subblocks. During hierarchical block matching, the bilinear window is used as an OBM window, where the size of extended window is 16×16. Given initial disparity field, pixelwise disparity estimation is performed using the cost function derived based on MRF/GRF framework. The resulting disparity field is mapped to the range $(0 - 255)$ to show the result clearly. Figure 4 compares the results of various disparity estimation scheme to show the effectiveness of the proposed framework.



(a) (b)

Figure 2: Test Images (Lab.pgm). (a) right (target) image (b) left (reference) image. The size of the image is 320×240.



(a) (b)

Figure 3: Intensity Edge and Segmentation Map (Lab.pgm). (a) Intensity edge (b) Initial segmentation map.

As shown in Figure 4 (a), block matching provide good initial disparity since it helps to avoid local minima. However, the resulting disparity sacrifies details along object boundary. Meanwhile the resulting disparity field using pixelwise estimation provide details but suffers from noise effects as shown in Figure 4 (b). A good tradeoff between block and pixel is block segmentation approach but it also suffers from local minima due to various noise effects including occlusion effects as shown in 4 (c). As shown in 4 (d), the proposed estimation scheme yields relatively smooth yet sharp disparity field by properly exploiting edge and neighboring, as well as initial, disparity information.

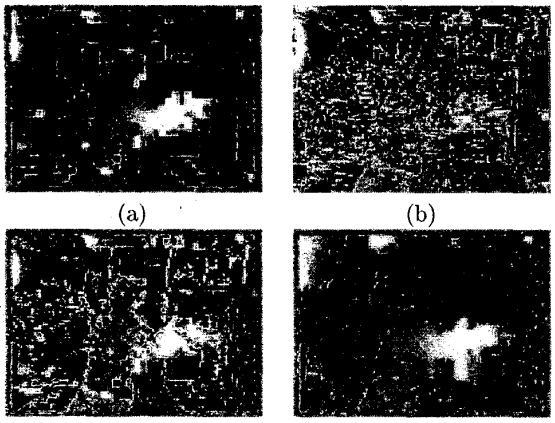


Figure 4: Comparison of Disparity Estimation (Lab.pgm). (a) Fixed size block matching (b) Pixelwise matching (c) Variable size block matching (d) Proposed hybrid matching. The initial block size is 8×8 and search window is $(0, 32)$.

Figure 5 (a) and (b) show the resulting OOI and the corresponding disparity information. The main assumption is that the OOI can be characterized by depth and edge information. As shown, the proposed scheme results in clear segmentation boundary yet smooth depth field.

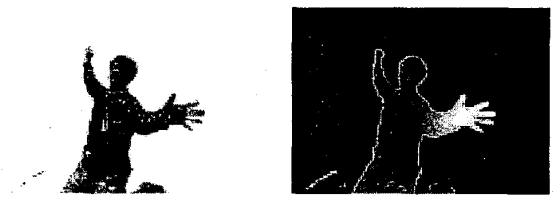


Figure 5: Image segmentation using disparity and intensity edges (Lab.pgm). (a) Segmented objects of interest (b) Disparity of the segmented OOI

In this paper we propose a coherent object segmentation scheme using stereo images, which is an important step in z-keying for the photo-realistic mixed reality. Even in mono sequence transmission scenario, the proposed segmentation scheme is useful because the cost of stereo-based segmentation may be cheaper than those of the other segmentation schemes at the encoder.

As shown in 5 (f), the proposed scheme achieves reasonable segmentation results by properly combining smooth disparity field and edge information. It is important step forwarding to achieve object-based functionalities in interactive multimedia applications, such as object-based query, editing, manipulation, etc. Note also that object-based coding is more suitable for upcoming multimedia applications requiring object-based manipulation or transmission because contours or object boundaries provide compact representation and convey much of the semantic content [1, 17]. In addition, the reconstructed image less suffers from the visual artifacts along the object boundaries.

Note however that perfect object segmentation is almost impossible without semantic knowledge about the scene. In addition various noise sources due to limited camera characteristics or continuously changing lighting condition as well as occlusion obstruct successful segmentation. As a result, as shown, sometimes resulting contours do not always match with real boundaries and the smooth estimation fails along occlusion areas, which may make annoying visual effects when we merge the objects into another image using z-keying. In other to solve this occlusion problem we are considering using one more camera, *e.g.* tri-camera system. Another unsolved issue is the quantitative measurement of the performance of segmentation schemes.

References

[1] M. Kunt, A. Ikonomopoulos, and M. Kocher, "Second-generation image coding techniques," *Proc. of the IEEE*, vol. 73, no. 4, pp. 549–574, Apr. 1985.

[2] M Chang, M Tekalp, and I. Sezan, "Simultaneous motion estimation and segmentation," *IEEE Trans. on IP*, vol. 6, pp. 1326–1333, Sept. 1997.

[3] R. Castagno, T. Ebrahimi, and M. Kunt, "Video segmentation based on multiple features for interactive multimedia applications," *IEEE Trans. on CSVT*, vol. 8, no. 5, pp. 526–571, Sept. 1998.

[4] E. Francois and B. Chupeau, "Depth-based segmentation," *IEEE Trans. on CSVT*, pp. 237–239, Feb. 1997.

[5] E. Izquierdo, "Image analysis for 3d modeling, rendering and virtual view generation," *CVIU*, vol. 71, no. 2, pp. 231–253, 1998.

[6] W. Woo and A. Ortega, "Stereo image compression based on the disparity compensation using the MRF model," in *Proc. SPIE VCIP*, Mar. 1996, vol. 2727, pp. 28–41.

[7] W. Woo and A. Ortega, "Stereo image compression based on the disparity field segmentation," in *Proc. SPIE EI-VCIP*, Feb. 1997, vol. 3024, pp. 391–402.

[8] W. Woo and A. Ortega, "Modified overlapped block matching for stereo image coding," in *Proc. SPIE EI-VCIP*, Jan. 1999, vol. 3653.

[9] W. Woo and Y. Iwadate, "Object-oriented hybrid image segmentation using stereo images," in *Proc. CVIP'00*, Jan. 2000.

[10] J.E. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. Royal Statistical. Soc.*, vol. B36, pp. 192–236, 1974.

[11] C. Chu and K. Aggarawal, "The integration of image segmentation maps using region and edge information," *IEEE Trans. on PAMI*, vol. 15, no. 12, pp. 1241–1252, Dec. 1993.

[12] W. Woo and A. Ortega, "Overlapped block disparity compensation with adaptive windows for stereo image coding," *IEEE Trans. on CSVT*, to appear, Mar. 2000.

[13] M. Levine, D. O'Handly, and G. Yagi, "Computer determination of depth map," *CGIP*, vol. 2, no. 4, pp. 131–150, 1973.

[14] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, pp. 283–287, 1976.

[15] E. Izquierdo, "Disparity/segmentation analysis: Matching with an adaptive window and depth-driven segmentation," *IEEE Trans. on CSVT*, vol. 9, no. 4, pp. 589–607, June 1999.

[16] B.K.P. Horn, *Robot Vision*, The MIT Press, 1986.

[17] P.J.L van Beek, *Edge-based Image Representation and Coding*, Thesis Technische University Delft, 1995.