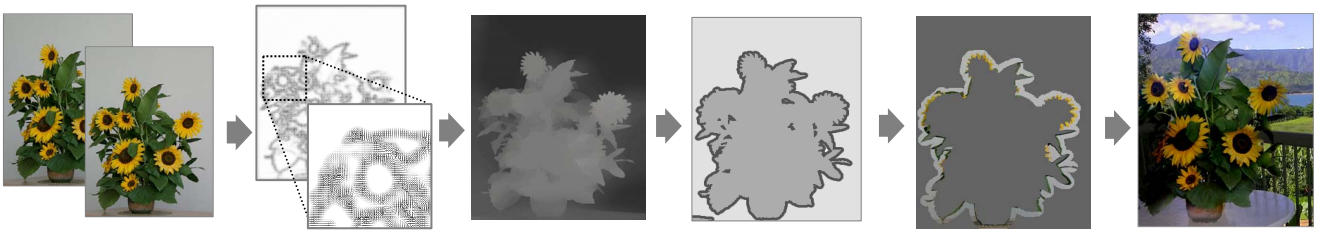
DEPTH DIFFUSION OBJECTS (DEDIO) - A SEAMLESS OBJECT-BASED APPROACH FOR TV APPLICATIONS

Jangheon Kim, Matthias Kunter, and Thomas Sikora
Department of Communication Systems, Technical University of Berlin,
Einsteinufer 17, 10587 Berlin, Germany
E-mail : {j.kim, kunter, sikora}@nue.tu-berlin.de



(a) stereoscopic images (b) edge flow field (c) depth diffusion (c) depth-based trimap (d) membrane harmonizing (e) new background

Figure 1. Procedure of "*DEDIO*"

ABSTRACT

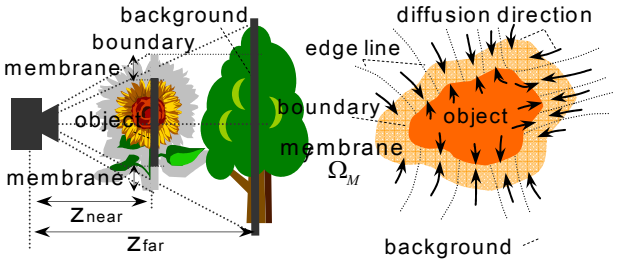
This paper proposes a novel and seamless object-based method for visual contents fusion called "*Depth Diffusion Object (DEDIO)*". It employs depth-based object segmentation in multi-view scenes and allows the composition of new scenes conveying a natural image impression. Using spatial anisotropic diffusion, the homogeneous regions of a scene which have similar depth are smoothly regularized following the spatial variation while discontinuities are preserved. The object's shape is automatically extracted by the depth range. In order to fuse the object with new background images we use membrane harmonizing applying directional constrained diffusion to remove visual seams and blend object, i.e. the remaining background area in the segmented object near the shape boundary. Our system automatically maintains the seamless object composition quality, even if the object is merged with new scenes, having different lightening conditions.

1. INTRODUCTION

Object- or content-based schemes play a significant role for the interaction capability of audio-visual media and TV contents [1]. They have widely been researched to provide a more accurate video representation in low bit-rate video compression and support content-based functionalities such as object manipulation. In the developed MPEG-4 standard, video coding is performed for each object unit known as "Video Object Plane (VOP)". Texture, motion, depth and auxiliary data, etc. can be separately coded in bit streams using shape information [2]. Thus, the coding applications efficiently process each object with different quality and amount of data according to the requirements. However, video object segmentation, which aims to extract exactly meaningful objects from a background, is still one of the most difficult problems, even though it has been researched for more than thirty years. Especially the segmentation quality needs to be improved for the breakthrough of object- or content-based schemes.

A lot of semi-automatic methods have been proposed using user-defined rules, e.g. feature-based, contour-based or region-based which directly point out initial object [3]. The general process is to collect only meaningful groups of pixels which can be extracted and tracked. The tracking mechanisms evolve and propagate the initial object using boundary errors. However, semi-automatic methods have a serious disadvantage, the necessity of user initialization. Thus, they are not suitable for TV applications.

Fully automatic methods [4, 5, 6] apply the extraction rules based on specific characteristics of the scene or on known *a priori* information. For example, the face segmentation uses spectral characteristics of skin-color region in CIE chromaticity diagram [4]. Chroma-keying, i.e. the blue screen method [5], accurately eliminates the uniformly colored background. However it requires a specific set-up of the background and special care for lighting conditions to avoid shadows and reflecting blue light and to maintain uniform intensity. To overcome some limitation of the chroma-keying, Z-keying, i.e. a depth keying method [5] which uses distance-based measures, was proposed. A depth map can be accurately computed by cameras with depth sensors that supplement every video frame with an additional frame, the "Z-buffer". However, the method is restricted to indoor applications since the depth range of the sensor is limited. For both, indoor and outdoor conditions, depth estimation based on the correspondence between multiple cameras substitutes the use of depth sensors. However, the depth estimation [6] provides low spatial accuracy in the segmentation results when the objects are sparsely textured. If a segmented object is laid over new background, old background areas which are extracted with the object can be distinctly recognized as "seam" in the object boundary.



(a) depth-trimap generation (b) membrane harmonizing

Figure 2. Setup of depth trimap and membrane harmonizing

Thus, partial replacement of the visible seam is required for pixel-wise accuracy. The precise definitions of the object boundary and coherence can be used to maintain the spatial accuracy.

In this paper we propose a new fully automatic object-based method which exploits spatial anisotropic diffusion for the depth field estimation. Figure 1 shows the procedure of the method. First, a full resolution depth field from multiple camera views is obtained by dense disparity estimation using spatial anisotropic diffusion. Since for spatial anisotropic diffusion the diffusion coefficient decreases at edges with steep intensity gradients, boundary over-smoothing can be avoided. The object can be extracted using the depth homogeneity since the depth inside the object is coarsely smoothed. A trimap is set up from the depth discontinuity as Figure 2 illustrates. After merging the object with the new image we apply diffusion directional blending from the new background into the membrane area of the object. By suitably incorporating the edge-preserved depth and the diffusive flow direction, blending direction and the stop line of blending are exactly localized. The membrane of the object is naturally blended with the new background but the important edges in object can be preserved. Consequently, visible seams due to the old background are entirely assimilated in the new background.

2. RELATED WORKS

Traditional object-based methods separate foreground objects from a video by shape information. In the image synthesis the foreground is simply overlapped with the background. This approach cannot assure natural and seamless image quality. Our method is more closely related with seamless composition (or matting) methods which blend the extracted object into new scenes. Seamless composition techniques have been researched for image (or video) editing, photography matting and film production. They are usually based on the manual or semi-automatic processes which need user's assistance.

The most natural image composition approaches are based on two different methods, i.e. the Bayesian matting method [7, 8] and the Poisson matting method [9, 10]. Bayesian matting methods require a user-defined shape map to initialize a trimap. The background and foreground color distribution are assumed to be mixtures of Gaussians. Fractional blending of foreground and background colors is conducted using a maximum-likelihood criterion. An extension of this method is also applied for video sequences [8]. The user draws an initial trimaps at specific key frames, which are interpolated across the video using forward and backward optical flow. The Poisson matting method efficiently solves the boundary condition problem in variational fields. A PDE-based solution iteratively devises the continuation of the flow direction and stops at the boundary of the selected region. The techniques fairly bridges narrow gaps in relative texture-free regions, information loss can be recovered and inpainted [12]. Poisson matting [10] solves the seamless composition problem by assuming that foreground and background are slowly varying. Although these methods result in a very natural composition quality, they need manual processes for large video sequences, far from real-time implementation for audio-visual media and TV contents.

Unassisted natural video matting systems are proposed in [6] and [11]. McGuire [11] solves the fully dynamic video matting problem using a trimap segmentation which is obtained from defocused images. The solution is directly driveded in filter-based image formation equations. Unfortunately this method requires stronger illumination than normal cameras since the sensors receive only small amount of incident light. Very similar to our method is the depth-based video proposed in [6]. The authors capture video sequences with a horizontal array of eight cameras placed over about two meters. They compute depth from stereo disparity using sophisticated region processing, and construct a trimap from depth discrepancies. The image/ video composition is performed using the Bayesian matting method. However, the results show only compositions of the object at the same position with the same background. The composition quality with new background and the boundary preservation quality for the object are not given.

3. DEPTH DIFFUSION OBJECT (DEDIO)

3.1. Edge flow dense depth disparity estimation

The proposed method segments the foreground object on 3D data of the FOV (field of view), i.e. both the color data and the depth data for each pixel. The depth $z=f\cdot(b/d)$ is estimated using disparity vector d between corresponding intensity values of multiple cameras on baseline b which have the focal length f. The different disparities allow creating parallel multi-layered depth objects, each being set at some specific depths. Unfortunately, depth estimation has to deal with several difficulties, e.g. the ambiguity of local image structure due to image noise, unbalanced brightness, similar texture and occlusion. If two pixels in the same image look alike, it may be impossible to find corresponding pixels in the other image based on only local appearance. To solve the problem, a supporting region is established by analyzing the local structure. Then, the depth is estimated in constraint of the support region. We define the support region by the edge flow scheme [13] employing a smoothly varying oriented structure as

$$\eta = \theta_+ = \frac{\nabla I_\sigma}{\|\nabla I_\sigma\|} \quad \text{and} \quad \xi = \theta_- = \frac{\nabla I_\sigma}{\|\nabla I_\sigma\|}^{\perp} \tag{1}$$

The unit vectors η and ξ are defined by gradient direction of the image Gaussian and its orthogonal direction, i.e. the isophote direction, respectively. The two variation orientation θ_+ and θ_- correspond to the vector edges of gradient and isophote. The scale parameter σ of the Gaussian-filter kernel is used to control the boundary strength.

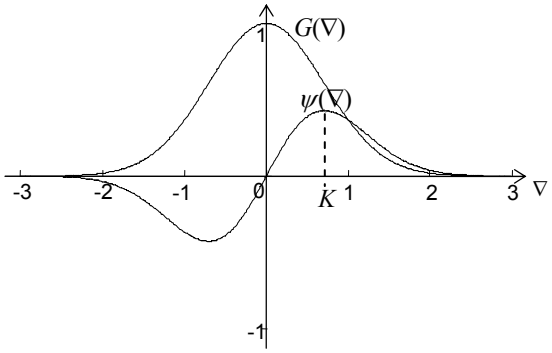


Figure 3. Anisotropic diffusion function

The Gaussian pre-filtering avoids the detection and the emphasis of initial noise and provides the best trade-off between the detection and localization performance.

$$I_\sigma(x,y) = (2\pi\sigma^2)^{-1} e^{-(x^2+y^2)/(2\sigma^2)} * I(x,y) \tag{2}$$

According to the *Helmholtz* theory, any vector field $\vec{F}$ can be represented as a sum of a conservative and solenoidal vector field with vector potential $\vec{A}$.

$$\vec{F} = \vec{F}_{con} + \vec{F}_{sol} = -\vec{\nabla} V + \vec{\nabla} \times \vec{A} \tag{3}$$

Taking the divergence of both side as

$$\vec{\nabla} \cdot \vec{F} = -\Delta V + \vec{\nabla} \cdot (\vec{\nabla} \times \vec{A}) \tag{4}$$

where Δ is the *Laplacian*. Since the second term is zero, the edge flow function V can be solved by the Poisson equation. Let the image be a continuous function which is only divided by edges into $N+1$ regions $\{R_0, \ldots, R_N\}$. The edge is calculated by a cost function as

$$E(V) = \sum_i \int_{R_i} \int_{R_i} w_\sigma(p_1, p_2) dp_1 dp_2 \tag{5}$$

where $w_\sigma(p_1,p_2) = \| p_1 - p_2 \|$ is a positive, symmetric dissimilarity function between the neighboring pixels of $(p_1[(\theta_+, \theta_-), \sigma], p_2[(\theta_+, \theta_-), \sigma]) \in R$ on the smoothly varying oriented structure of (1). The support region W is obtained by combining similar regions. The cost function between i-th and j-th region is

$$E(W(V)) = \int_{R_i} \int_{R_j} w_\sigma(p_1, p_2) dp_1 dp_2 + E(V) \tag{6}$$

Dense disparity vectors d are locally estimated with the support region W enclosed in the boundary and refined for the scale σ. The final energy function then yields in

$$E_\Omega(d) = \int_\Omega [I_{l,\sigma}(x,y) - I_{r,\sigma}(x + d_{l\to r}(x) \in W(V), y)]^2 dxdy + \lambda \int_\Omega e_\sigma(d) dxdy \tag{7}$$

Ω is the image domain, the subscripts denote the matching direction, e.g. $l \to r$ for left-to-right direction and e_σ is a regularization term with the Lagrange multiplier λ. The enclosed region is used as the support region for the matching. This method results in a more accurate local disparity estimation solution due to restricted matching errors.

3.2. Regularization with spatial anisotropic diffusion

We globally regularize the locally estimated disparity vector field which is obtained by minimizing (7) applying an edge-preserving spatial anisotropic diffusion.

$$e_\sigma = G(\|\nabla I_{l,\sigma}\|) \nabla d_{l\to r} \tag{8}$$

This is a modified version for disparity of the discrete Perona and Malik model [14] which has the form $\Psi(\nabla) = G(\nabla)\nabla$ ' as flux function. $G(\nabla)$ is an anisotropic diffusion function which is called "*edge-stopping function*" – used to modify the diffusion coefficient at edges and to derive discontinuities. A suitable choice of G is

$$G(\nabla) = e^{-(\nabla^2 / K^2)} \tag{9}$$

where a positive constant K controls the level of contrast of edges affecting the diffusion process as Figure 3 shows. We solve the energy-minimization problem in (7) by discretizing the following equation using finite differences.

$$\begin{aligned} \frac{d_{l\to r}^{t+1} - d_{l\to r}^{t}}{\tau_\sigma} = \lambda \mathrm{div} \left[\left(\frac{G(\|\nabla I_{l,\sigma}(x,y)\|)}{\|\nabla I_{l,\sigma}(x,y)\|} \right) \nabla d_{l\to r}^{t}(x,y) \right] + \\ \frac{\partial I_{r,\sigma}(x + d_{l\to r}^{t}(x), y)}{\partial x} \left[\left[I_{l,\sigma}(x,y) - I_{r,\sigma}(x + d_{l\to r}(x), y) \in W(V) \right] \right. \\ \left. + \frac{\partial I_{r,\sigma}(x + d_{l\to r}^{t}(x), y)}{\partial x} (d_{l\to r}^{t+1} - d_{l\to r}^{t}) \right] \end{aligned} \tag{10}$$

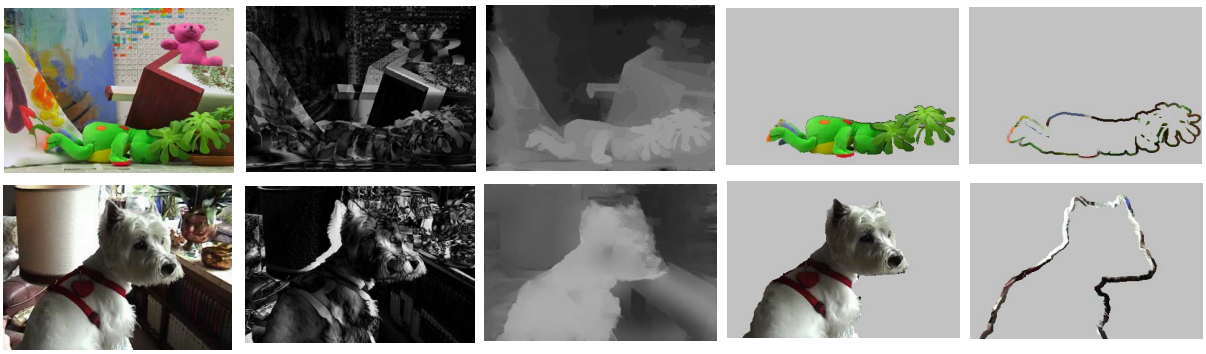
An inhomogeneous time diffusion process with discrete sampling solves the problem. Different pixels diffuse at a different time scale related to the pixel confidence. By increasing the time step τ_σ while refining resolution with the scale σ, higher confidence pixels diffuse much slower than low confidence pixels. The details of the solution are given in [15]. In the coarse scale, the disparity estimation using diffusion is performed for a wider range between strong edges following the smoothed field. With further iterations, this gradually occurs in a narrower range between weaker edges as well. Thus, falling into local minima of true-backward diffusion and the over-split problem in large gradient areas, e.g. shadows and texture regions, etc. can efficiently be avoided.

3.3. Membrane harmonizing with diffusion constraint

We extract the objects and generate its membrane Ω_M from the base image I_l using a threshold TH of the variation of disparity $\nabla d_{l\to r}$. The membrane is set up at the position of the object boundary by expanding the boundary into the foreground and the background region.

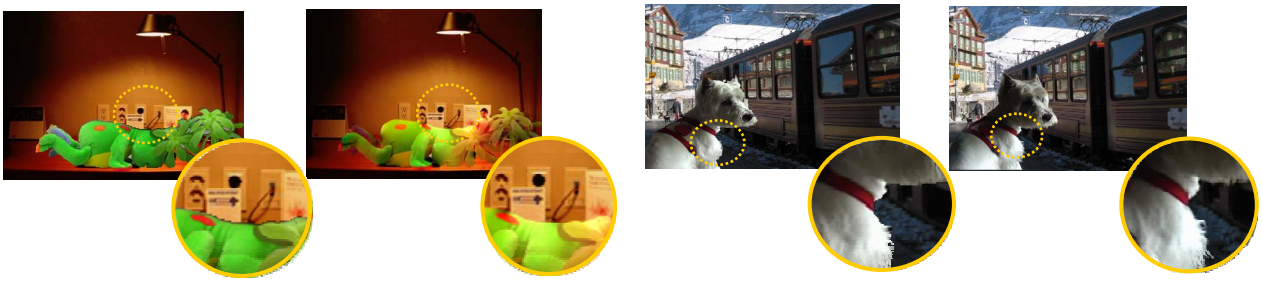
$$\text{Set } \Omega_M \text{ for } \Omega \in \nabla d_{l\to r} > TH \tag{11}$$

While anisotropic diffusion enhances depth homogeneity and preserves the discontinuity $\partial\Omega$, boundary pixels should be harmonized with new background. Strong variations on the object boundary should be preserved for conserving high-frequency details e.g. hair, wool and fuzz, etc. With *Dirichlet* boundary condition (it specifies the *Laplacian* of an unknown function over the domain of interest, along with the unknown functional values over the boundary of the object), we solve the two-fold blending problem for object Ω with a membrane $\Omega_M|_{\partial\Omega}$.



(a) one viewpoint (b) difference with another view (c) dense depth map (d) keying object (e) membrane domain

Figure 4. Simulation results of depth-trimap (The upper images are for "*Teddy*" and the below images are for "*JoJo*")



(a) *"Teddy"* (b) *"JoJo"*

Figure 5. Efficiency with new background (the left and right images in (a) and (b) respectively show MPEG-4 VOP and *DEDIO*)

For the pixels in the membrane, the seams of the object are blended and harmonized with the new background color $I_{new}(p)$ using diffusion direction and its power. By substituting $\nabla d_{l\to r}$ in (8) into $\nabla I_{new,}$, we blend the object with new background with the smoothly-varying directional constraint and power of scales.

$$G(\|\nabla I_{l,\sigma}\|)\nabla I_{new} = 0 \quad \text{for all} \quad p \in \{\Omega, \Omega_M\}|_{\partial\Omega} \tag{12}$$

4. RESULTS AND CONCLUSION

We first show the results for depth-trimap generation in Figure 4. Since our method efficiently yields very smooth depth for each object (Figure 4c), the object can easily be extracted (Figure 4d). The discontinuity preserving anisotropic regularization supports a good localization performance for depth-based segmentation. At the depth discontinuity, the membrane, which is important for blending the object boundary with new background pixels, is shown in Figure 4e.

The visual seams at the object boundary can be efficiently removed. In Figure 5, we compare the naturalness of the image perception for different background scenes between the MPEG-4 VOP and DEDIO. Note, that for results of MPEG-4 VOP we also generate the object by our reliable anisotropic depth diffusion approach. In Figure 5a, we challenge brightness adaptation by fusing object "*Teddy*" with spot lighting. In the case of MPEG-4 VOP, seams, i.e. the darker pixels at the boundary, can be perceived although segmentation is very accurate. In contrast DEDIO generates a natural composition adapting excellently to the change of the lighting condition. In Figure 5b, the membrane harmonizing of our method conserves the high-frequency details of the object boundary e.g. the fur at the boundary of dog "*JoJo*". Therefore, DEDIO can be utilized for high-quality audio-visual interactive contents for future TV.

6. REFERENCES

[1] A. Smolić, P. Kauff, "Interactive 3D Video Representation and Coding Technologies", in Proc. of the *IEEE, Special Issue on Advances in Video Coding and Delivery*, Vol. 93, No. 1, 2005.

[2] ISO/IEC 14496-2 Final Committee Draft, *Information Technology - Coding of Audio-Visual Objects: Visual*, May 1998.

[3] A. Cavallaro, T. Ebrahimi, "Object-based video: extraction tools, evaluation metrics and applications", in Proc. of *SPIE VCIP*, 2003.

[4] L. Lucchese, S. Mitra, "Color image segmentation: A State-of-the-Art survey", in Proc. of *the Indian National Science Academy*, 2001.

[5] R. Gvili, A. Kaplan, E. Ofek, G. Yahav, "Depth keying", in Proc. of *SPIE Stereoscopic Displays and Virtual Reality Systems*, 2003.

[6] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, R. Szeliski, "High-quality video view interpolation using a layered representation", *ACM Trans. on Graphics 2004.*

[7] Y. Chuang, B. Curless, D. Salesin, R. Szeliski, "A Bayesian approach to digital matting," In Proc. of *IEEE CVPR 2001.*

[8] Y. Chuang, A. Agarwala, B. Curless, D. Salesin, R. Szeliski, "Video matting of complex scenes", *ACM Trans. on Graphics*, pp. 243–248, 2002.

[9] P. Pérez, M. Gangnet, A. Blake, "Poisson image editing", *ACM Trans. on Graphics*, pp. 313–318, 2003.

[10] J. Sun, J. Jia, C. Tang, H. Shum, "Poisson Matting", *ACM Trans. on Graphics,* pp. 315–321, 2004.

[11] M. McGuire, W. Matusik, H. Pfister, J. Hughes, F.Durand, "Defocus Video Matting," *ACM Trans. on Graphics,* 2005.

[12] M. Bertalmío, G. Sapiro, V. Caselles and C. Ballester, "Image Inpainting", In Proc. of *ACM SIGGRAPH 2000.*

[13] B. Sumengen, B. Manjunath, "Multi-scale Edge Detection and Image Segmentation," In Proc. of *EUSIPCO*, 2005.

[14] P. Perona, J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. on PAMI* 12(7), pp. 629-639, 1990.

[15] J. Kim, T. Sikora, "Gaussian Scale-Space Dense Disparity Estimation with Anisotropic Disparity-Field Diffusion," In Proc. of *IEEE 3DIM*, 2005.