특집논문-09-14-4-07

다중 단계 신호의 적응적 전파를 통한 동일 장면 영상의 이원 영역화

이 수 찬^{a)*}, 윤 일 동^{b)}, 이 상 옥^{c)}

Bilayer Segmentation of Consistent Scene Images by Propagation of Multi-level Cues with Adaptive Confidence

Soochahn Lee^{a)*}, Il Dong Yun^{b)}, and Sang Uk Lee^{c)}

요 약

최근까지 단일 영상이나 동영상 영역화하는 기법들은 다양하게 제시되어 왔으나, 유사한 장면에 대한 여러 장의 영상을 동시에 영역화하는 기법은 많지 않았다. 본 논문에서는 한 장소에서 연속적으로 촬영하였거나 전경 물체가 유사한 여러 영상들을 동일 장면 영상으로 정의하고, 이런 동일 장면 영상들을 적은 양의 사용자 입력을 통해 효과적으로 영역화하는 기법을 제안한다. 구체적으로, 사용자가 최초의 영상 한 장을 직접 영역화한 후, 그 영상의 영역화 결과와 영상의 특성을 토대로 다중 단계 신호를 적응적 가중치를 주어서 인접 영상으로 전파하고, 이를 통해 제안하는 기법은 인접 영상을 반복적으로 영역화한다. 영역화는 마르코프 랜덤 장에서의 에너지 최소화를 통해 이루어지는데, 전파되는 신호는 각 픽셀에 대한 에너지를 정의하는 바탕이 되며, 픽셀, 픽셀 패치, 그리고 영상 전체로부터 비롯되었는가에 따라 낮은 단계, 중간 단계, 그리고 높은 단계의 신호로 지칭된다. 또한 에너지 최소화 틀 안에서 전파된 신호를 통해 정의되는 에너지 역시 낮은 단계, 중간 단계, 그리고 높은 단계의 세 단계로 정의한다. 이런 과정을 통해 전파된 신호를 최대한 다양하게 활용하고, 이를 통해 다양한 영상에 영역화 결과가 일관되게 유지된다. 다양한 동일 장면 영상들에 제안하는 기법을 적용하여 성능을 평가하고, 픽셀 패치를 바탕으로 하는 중간 단계 신호만을 이용한 결과와 제안하는 다중 신호를 적용하는 기법의 결과를 비교한다.

Abstract

So far, many methods for segmenting single images or video have been proposed, but few methods have dealt with multiple images with analogous content. These images, which we term consistent scene images, include concurrent images of a scene and gathered images of a similar foreground, and may be collectively utilized to describe a scene or as input images for multi-view stereo. In this paper, we present a method to segment these images with minimum user input, specifically, manual segmentation of one image, by iteratively propagating information via multi-level cues with adaptive confidence depending on the nature of the images. Propagated cues are used as the bases to compute multi-level potentials in an MRF framework, and segmentation is done by energy minimization. Both cues and potentials are classified as low-, mid-, and high- levels based on whether they pertain to pixels, patches, and shapes. A major aspect of our approach is utilizing mid-level cues to compute low- and mid- level potentials, and high-level cues to compute low-, mid-, and high- level potentials, thereby making use of inherent information. Through this process, the proposed method attempts to maximize the amount of both extracted and utilized information in order to maximize the consistency of the segmentation. We demonstrate the effectiveness of the proposed method on several sets of consistent scene images and provide a comparison with results based only on mid-level cues [1].

Keywords : Image Set Segmentation, Consistent Scene Images, Multi-level Cues, Adaptive Confidence

I. Introduction

The problem of segmenting images into regions with common properties has long been an actively researched topic in computer vision and graphics with applications ranging from 3-D reconstruction and object recognition to image synthesis and virtual reality. In many cases the need to segment multiple images with analogous content arise and segmenting each image separately requires repetitive user input. Although numerous segmentation methods dealing with various data types have been proposed, this problem has yet to be properly addressed. This paper presents a robust method to segment these images which we call consistent scene images. Formally, consistent scene images are defined as images with consistent appearance, be it the object of interest, the foreground, or the background, with limited but general variation between each image.

The segmentation problem has a broad spectrum and can be classified in many ways depending on the type of data, i.e. single image, multiple images, or video, whether user supervision or higher level cue is provided, and the number of output segments. Needless to say, single image segmentation is by far the most researched. Unsupervised methods that cluster pixels with homogeneous properties such as color or intensity into small scale superpixels[2,3], or large scale regions[4] tend to create multiple segments. Many times these segments tend to contradict human intuition by

dividing pixels that belong to a single object, but play an important part as the building blocks for more elaborate segmentations methods. For segmentation results to follow human intuition, some sort of additional information must be provided, resulting in supervised methods. The information may be confined to user input seeds[5,6], or expanded to high level knowledge of the object to segment [7,8]. These methods usually segment the image into bi-layers, i.e. the foreground and the background. Segmentation based on graph cuts[5] have especially played an important part not only in providing a simple user guided segmentation method, but also as the basic framework of energy minimization in a pairwise MRF by finding the min-cut of a graph, where multiple levels of cues can be incorporated.

Based on the proliferation of effective single image segmentation methods, many methods for segmenting video data have consequently been proposed. Based on superpixels from mean shift [2] as building blocks and applying graph cuts [5] in a 3-D space with a practical user seed input framework, the methods presented in [9] and [10] have produced convincing results. Also, real-time methods that segment the input video stream from a web camera have also been proposed. These methods are characterized by the cues they utilize, such as stereo [11], stationary background [12], and motion parameters [13,14].

Although many effective methods for segmenting single images and video have been proposed, the need to segment multiple images containing consistent content with unknown variations creates a new problem. For instance, if the user tries to reconstruct the 3D structure of an object from images taken at various time and place such as the images in Fig 1, the object must first be segmented since camera calibration techniques cannot be directly applied due to the change of background. Video segmentation will produce fruitless results, and repeated use of single image segmentation methods will most likely cause a burden on

a) 서울대학교 자동화시스템연구소

Automation and Systems Research Institute, Seoul National University

b) 한국외국어대학교 용인캠퍼스 디지털정보공학과

School of Digital Information Engineering, Hankuk University of Foreign Studies

c) 서울대학교 전기·컴퓨터공학부

School of Electrical Engineering and Computer Science, Seoul National University

✉ 교신저자 : 이수찬(redhouse@cvi.snu.kr)

※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (20090053539).

· 접수일(2009년3월20일), 수정일(2009년7월28일), 게재확정일(2009년7월28일)

그림 1. VOC2006 데이터베이스 [16]의 영상들 중 선택된 BUS 집합의 원 영상 (윗 줄) 및 영역화된 전경 (아래 줄). 영상들은 각기 다른 때와 장소에서 촬영한 2층 버스 영상들로 이루어짐

Fig. 1. Images of the BUS set comprising of image of a two story bus taken at multiple times and places collected from the VOC2006 database [16] and segmentation results obtained from the method proposed in the paper

the user. Another such instance may arise when the user wants to synthesize a background for multiple images of a particular scene. Therefore, the problem of segmenting consistent scene images must deal with the problem of maximizing the effect of minimal user input by utilizing every possible aspect of user provided information to create a robust segmentation routine for multiple images with consistent content.

In this paper we deal with this problem by propagating multiple cues with adaptive confidence from an image segmented by the user. By utilizing multiple cues and evaluating their confidence, we are able to maximize the propagated information initially provided by the user, thereby maximizing the robustness and consistency of segmentation for all images. The utilized cues can be classified as low, mid, and high level cues depending on their domain, and their confidences are evaluated based on the nature of the relation between the segmented image and the image to segment. In the process of propagation, the foreground motion between the images is also estimated to maximize the correlation of the cues. The propagated cues and their con-

fidence are utilized in an MRF based bilayer segmentation framework as the bases for computing multi-level potentials of pixels, patches, and boundaries. Energy minimization is performed by finding the min-cut [15] of the MRF, and cue propagation/segmentation is iterated until all the images have been segmented.

The problem of cosegmentation [17] which segments a pair of images simultaneously based on the common content, and multi-view segmentation [18] which segments images of a certain object taken at various viewpoints are closely related to the problem at hand, but not without several differences. Cosegmentation automatically segments pairs of images with common foregrounds and contradictory backgrounds. In the case an image has been segmented, only its foreground and background histograms are considered to segment the other. For multi-view segmentation, the images must be calibrated and a sufficient number of images must be provided so that the images encompass the whole object of interest. Also, the foreground object must be stationary. In our problem, the number of consistent scene images may be more than two with arbitrary

viewpoints and variation in appearance and uncalibrated. We note that although the use of multiple cues was inspired by the method of [19], it deals with learning the cues from multiple images, where as we directly extract cues from a single segmented image.

The rest of the paper is organized as follows. In section 2, we present a general overview of the proposed method. In section 3 and section 4, we explain the details of determining multi-level cues and their confidences and propagating them for iterative segmentation, respectively. Experimental results are presented in section 5, and the paper is concluded in section 6.

II. Overview of the Proposed Method

The proposed method to segment consistent scene images is initiated based on a user segmented image. From the segmented image, we extract low, mid, and high level cues, and propagate them to the image we wish to segment. The confidence of each cue is computed by comparing the similarity of the segmented image I_{seg} and the image to segment I .

Multi-level cues are classified into low-level cue Q_l which is based on pixels, mid-level cue Q_m which is based on patches, and high-level cue Q_h which is based on the shape of the boundary. More specifically, the pixel distribution of the foreground I_{seg}^F and background I_{seg}^B of the segmented image constitute Q_l , while a dictionary of representative patches from I_{seg}^F and I_{seg}^B constitute Q_m . The shape of the boundary between I_{seg}^F and I_{seg}^B constitutes the high level cue, where the boundary of I is estimated by estimating the motion between I_{seg}^F and I denoted as T . The cues are categorized as the aforementioned three since Q_l represents the smallest scale of image components, Q_h represents the largest, and Q_m represent which are in

between.

The propagated cues are used as the criteria for segmenting I in an MRF energy minimization framework where many methods to integrate multiple cues such as [7,8,20] have been proposed. Also, image segmentation methods [5] and [6] based on the MRF have established favorable results. This framework involves computing potentials for multiple cliques on a random field $\mathbb{X}$ defined over a lattice $V=1,2,...,N$, including potentials for each random variable $x \in \mathbb{X}$ corresponding to pixel $p \in I$, pair of random variables (x_i, x_j) corresponding to pixel pair (p_i, p_j) , and vector of random variables $\mathbf{x} \in \mathbb{X}$ corresponding to patch p_θ . The segmentation is done by determining the configuration of $\mathbb{X}$ that minimizes the sum of all potentials which corresponds to a MAP estimation. Here, $(i,j) \in \{N\}$ denotes the set of neighboring lattice positions, $\theta \in \Theta$ denotes the set of all patches, and each x takes a value from the label set $L = \{F, B\}$. For a more detailed description of MRF based image labelling problems, please refer to [21].

Potentials are classified as pixel potentials, patch potentials, patch boundary potentials, and boundary potentials. We collectively term pixel potentials as low-level potentials Φ_l , patch potentials and patch boundary potentials as mid level potentials, Φ_m and boundary potentials as high level potentials Φ_h following the classification scheme used for cues. Based on this terminology we establish an important observation that potentials of equal or lower levels can be computed from each level of propagated cue. That is, low-level cue Q_l , mid-level cue Q_m , and high level cue Q_h provide bases to compute low-level potentials Φ_l , Φ_{l_m} , and Φ_h , Q_m and Q_h provide bases to compute mid-level potentials Φ_{m_m} and Φ_{m_h} , and Q_h provides a basis to compute the high-level potential Φ_{h_h} . For clarity of notation we distinguish between three types of patches, namely, super-pixels θ patches ω stored in the patch dictionaries, and boundary patches μ .

The potentials are combined through the energy equation:

$$E(\mathbb{X}) = \lambda_{\phi_l} \sum_{i \in V} \phi(x_i) + \sum_{i \in V, (i,j) \in \{N\}} \phi(x_i, x_j) + \lambda_m \sum_{\theta \in \Theta} \phi_{m_m}(x_\theta | Q_m) \quad (1)$$

where

$$\phi(x_i) = c_l \phi_l(x_i | Q_l) + c_m \phi_m(x_i | Q_m) + c_h \phi_h(x_i | Q_h) \quad (2)$$

and

$$\phi(x_i, x_j) = \lambda_{\phi_m} \phi_{m_h}(x_i, x_j | Q_h; c_{h_h}) + \lambda_{\phi_h} \phi_{h_h}(x_i, x_j | Q_h; c_{h_h}) + \lambda_b \phi_b(x_i, x_j). \quad (3)$$

Here, c_l , c_m , and c_h , $\bar{c}_h$, $c_{h,\mu}$ represent the confidence of Q_l , Q_m , and the average confidence of Q_h , over boundary patches μ , respectively. Also, ϕ_b represents the gradient based pairwise terms enforcing smooth boundaries, while λ_{ϕ_l} , λ_m , λ_{ϕ_m} , λ_{ϕ_h} , and λ_b represent the parameters that control the ratio between the potentials.

As mentioned, segmentation results for I are determined by minimizing E of (1). After segmentation, the newly segmented image I becomes I_{seg} , the next image $I_{\neq xt}$ becomes I , and the process is iterated until all the images have been segmented. During the iteration process, instead of replacing old cues with new ones, newly extracted cues are mixed in, in order to accumulate useful data needed for accurate segmentation.

III. Multi-Level Cues with Adaptive Confidence

The multi-level cues each offer a distinctive clue on how to segment I . Statistics of the overall distribution of the color of each pixel for both F and B provide a fundamental comparison between pixels in each region, while patches comprise a more specific description with less abstraction. On the other hand, the boundaries between F

and B provide the shape of each region and the details on the gradient properties that occur where the regions are divided. We attempt to maximize the propagated information by utilizing all these cues as well as compute their confidence in order to compensate for the inconsistency between I_{seg} and I .

Low-Level Cue: The distribution of the colors of pixels that constitute the foreground and background are approximated by 3-D color histograms. We construct two histograms H_F and H_B from pixels that belong to the foreground I_{seg}^F and background I_{seg}^B , respectively, of I_{seg} . Confidence of this cue is determined by comparing the distance between I_{seg} and I in the pixelwise color distribution domain. That is, we define the confidence c_l of Q_l from the distance D_l between histograms $H_{I_{seg}}$ and H_I of I_{seg} and I , respectively, through a negative exponential function $g(x; \sigma) = \exp(-x^2/\sigma^2)$. Therefore, the propagated low-level cue and its confidence are:

$$Q_l = \{H_{I_{seg}}^F, H_{I_{seg}}^B\}, \quad (4)$$

$$c_l = g(-D_l(H_{I_{seg}}, H_I); \sigma_{Q_l}) \quad (5)$$

Mid-Level Cue: We follow the philosophy proposed in [1] of first segmenting I_{seg} into superpixels $\Theta_{I_{seg}}$ and then extracting a constant ratio of patches from each $\theta_{I_{seg}}$, in order to correlate the number of extracted patches to the texture or homogeneity in I_{seg} . By extracting patches $\omega_{I_{seg}}$ from I_{seg}^F and I_{seg}^B , we create two dictionaries $\Omega_{I_{seg}}^F$ and $\Omega_{I_{seg}}^B$. Again, confidence is determined by comparing I_{seg} and I , but this time in their superpixel domain. More specifically, we create color signatures [22] $S_{I_{seg}}$ and S_I based on the superpixels for I_{seg} and I , respectively, and apply the distance D_m between $S_{I_{seg}}$ and S_I as the input to $g(x; \sigma)$. The propagated mid-level cue and its confidence are

$$Q_m = \{\Omega_{I_{seg}}^F, \Omega_{I_{seg}}^B\}, \quad (6)$$

$$c_m = g(-D_m(S_{I_{seg}}, S_I); \sigma_{Q_m}) \quad (7)$$

High-Level Cue: The boundary between the foreground and the background represent the outline of the foreground. This outline is the most important aspect of the shape in terms of bilayer segmentation, which is essentially the process of identifying the outline of the foreground. Therefore, we use the boundary between I_{seg}^F and I_{seg}^B as the high level cue. Since the relation between I_{seg}^F and the foreground of I must be estimated in order for the boundary of I_{seg} to be useful in estimating the boundary of I , we estimation the motion between through T . We apply the feature-based non-rigid registration method of [23] as the motion estimation scheme T and register I_{seg}^F into I , since it can handle various types of motion. The confidence

of the deformed boundary is based on how well the boundary coincides with the gradients or edges of I . Specifically, from the gradient map of I , we compute the gradient vector flow (GVF) [24] and compute the inner product between the normal of $T_I(I_{seg}^F)$ and the GVF at the boundary points to obtain $\bar{c}_h$. We apply the GVF since it represents distances as well as orientations of normals dispersed from nearby gradients. Unlike the c_l and c_m which are constants, $\bar{c}_h$ is a matrix with the same size as the image. The final propagated high level cue Q_h and its confidence are

$$Q_h = T_I(I_{seg}^F), \quad (8)$$

and

$$\bar{c}_h(i) = (\perp(\nabla I_{seg}(i)) \cdot \text{GVF}_I(i)) \times B(T_I(I_{seg}^F))(i) \quad (9)$$

그림 2. 추출된 다중 단계 정보 Q_l , Q_m , 그리고 Q_h 에 대한 설명. 사용자가 영역화한 영상 I_{seg} 으로부터 추출되어 새 영상 I 로 전파됨

Fig. 2. Description of the extracted multi-level cues Q_l , Q_m , and Q_h utilized for propagating information given by the user when segmenting I_{seg} to segment I

where

$$B(i) = \begin{cases} 1, & \text{if } i \text{ is on the boundary} \\ 0, & \text{otherwise} \end{cases}$$

and $\perp(v)$ represents the normal of vector v .

A visual description of each cue is given in Fig. 2.

IV. Iterative Segmentation from Propagated Cues

1. Computing Multi-Level Potentials

Based on the propagated cues we compute potentials of the pixels $p \in I$, patches $\overline{p_\theta} \in I$, and pixel pairs $(p_i, p_j) \in N$ where $(i, j) \in \{N\}$ for I . Following the classification scheme for cues, we term potentials based on pixels as low-level, patches as mid-level, and propagated boundary as high-level potentials. We derive multiple potentials from each level of cues to maximize the inherent information

and increase robustness. More specifically, we define three low level potentials $\Phi_l(I|Q_l)$, $\Phi_m(I|Q_m)$, and $\Phi_h(I|Q_h)$, two mid-level potentials $\Phi_{m_m}(I|Q_m)$ and $\Phi_{m_h}(I|Q_h; \overline{c_h})$, and one high-level potential $\Phi_h(I|Q_h; \overline{c_h})$. It can be seen that each level potential is computed from cues that are of same or higher levels. Fig. 3 gives an outline of each potential.

Low-level potentials: Computation of low-level potentials is straight forward in that they are computed from histograms H^F and H^B . The difference lies in how the histograms for each level are computed. For $\Phi_l(I|Q_l)$, the histograms have already been computed, resulting in $\Phi_l(I|\{H_{seg}^F, H_{seg}^B\})$. For $\Phi_m(I|Q_m)$, the histograms $H_{l,m}^F$, $H_{l,m}^B$ are computed by summing the foreground and background likelihoods $l^F(\theta_I|Q_m)$ and $l^B(\theta_I|Q_m)$ of the corresponding superpixel for all pixels. Here, the likelihood l_θ^F and l_θ^B for θ are determined when the mid-level potentials

그림 3. 영역화된 영상 I_{seg} 으로부터 추출된 정보를 바탕으로 I 를 나타내는 MRF에 적용되는 다중 단계 포텐셜에 대한 묘사

Fig. 3. Description of multi-level potentials for the MRF based on I computed from the cues propagated from I_{seg}

are computed and thus defined in section 4. For $\Phi_{l_h}(I|Q_h)$, histograms $H_{l_h}^F$, $H_{l_h}^B$ are computed from pixels corresponding to $T_I(I_{seg}^F)$ and $T_I(I_{seg}^B)$, respectively. From the respective histograms, foreground potentials are computed as

$$\phi_{l_h}(x_i = F | \{H_{l_h}^F, H_{l_h}^B\}) = -\log\left(\frac{H_{l_h}^F(p_i)}{H_{l_h}^F(p_i) + H_{l_h}^B(p_i)}\right), \quad (11)$$

$$\phi_{l_m}(x_i = F | \{H_{l_m}^F, H_{l_m}^B\}) = -\log\left(\frac{H_{l_m}^F(p_i)}{H_{l_m}^F(p_i) + H_{l_m}^B(p_i)}\right), \quad (12)$$

$$\phi_{l_b}(x_i = F | \{H_{l_b}^F, H_{l_b}^B\}) = -\log\left(\frac{H_{l_b}^F(p_i)}{H_{l_b}^F(p_i) + H_{l_b}^B(p_i)}\right), \quad (13)$$

where the value of p_i is the color and $H(p_i)$ is the value of the histogram corresponding to p_i . Background potentials are computed by replacing $H^B(p_i)$ in the numerator.

Mid-level potentials: Computation of $\Phi_{m_m}(I|Q_m)$ involves computing the likelihood $l(\theta)$ of each superpixel $\theta_I \in \Theta_I$. Again, we extract a certain ratio of patches $\Omega_I^{\theta_I}$ for each superpixel θ_I . For each $\omega_I \in \Omega_I^{\theta_I}$ we compute the distance to its k-nearest neighbor in Ω_I^F and use their average and variance [1] to obtain the foreground patch likelihood

$$l^F(\theta_I | Q_m) = \frac{g(\mathbb{E}[d_{knn}(\omega_I, \Omega_I^F)]; \sigma_d^2)}{g(\mathbb{E}[d_{knn}(\omega_I, \Omega_I^F)]; \sigma_d^2) + g(\mathbb{E}[d_{knn}(\omega_I, \Omega_I^B)]; \sigma_d^2)} \quad (14)$$

where $\mathbb{E}[\cdot]$ denotes average, $d_{knn}(\omega, \Omega)$ denotes the distance between ω and its k-nearest neighbor in Ω , and σ^2 denotes variance of the k-nn distances. The background likelihood is computed in the same manner, and $\Phi_{m_m}(I|Q_m)$ is computed as

$$\phi_{m_m}(x_{\theta_I} | \{\Omega_I^F, \Omega_I^B\}) = -|\theta_I| \log(l(\theta_I)). \quad (15)$$

where $|\theta|$ is the size of θ .

The patch boundary potential $\Phi_{m_h}(I|Q_h; \bar{c}_h)$ represents the potential for the boundary to be between pairs of variables (x_i, x_j) based on patches on the boundary of I_{seg} . The assumption is that the patches $\mu_{I_{seg}}$ on the boundary of I_{seg} and patches μ_I transformed through $T_I(I_{seg}^F)$ will have similar appearances. Therefore, we segment μ_I based on histograms $\{H_\mu^F, H_\mu^B\}$ computed from $\{\mu_{I_{seg}}^F, \mu_{I_{seg}}^B\}$, and estimate the boundary of I , namely B_m , as the average of the boundaries of all μ_I

$$B_m = \mathbb{E}_{\mu_I} [c_{h, \mu_I} | \mu_{I_{seg}}], \quad (17)$$

where c_{h, μ_I} is the average of c_h over μ_I . From the estimated boundary B_m , $\Phi_{m_h}(I|Q_h; c_h)$ is computed as

$$\phi_{m_h}(x_i, x_j | B_m) = \frac{\phi_{m_h}(x_i | B_m) + \phi_{m_h}(x_j | B_m)}{2} \quad (18)$$

where

$$\phi_{m_h}(x_i | B_m) = \frac{\sum_{\varepsilon \in B_m} w_\varepsilon(i) c_h(\varepsilon) d(i, \varepsilon)}{\rho_{B_m} \sum_{\varepsilon \in B_m} w_\varepsilon(i)}. \quad (19)$$

Here, $w_\varepsilon(i) = \exp\left(-\frac{d^2(i, \varepsilon)}{\rho^2}\right)$ and $\rho_{B_m} = \mathbb{E}_{\tau, \xi \in B_m} [d(\tau, \xi)]$

represent the weight of the influence of contour point ε and the average distance between points on the estimated boundary B_m , respectively. From (19) we are able to put more emphasis on estimated boundary points with higher confidence, and boundary points that are close to lattice point i .

High-level potential: The boundary potential $\Phi_{h_h}(I|Q_h;\overline{c_h})$ is computed in a similar manner to $\Phi_{m_h}(I|Q_h;\overline{c_h})$. The difference is that $B(T_I(I_{seg}^F))$ is directly applied without the boundary estimation process based on μ , thereby

$$\begin{aligned} & \phi_{h_h}(x_i, x_j | B(T_I(I_{seg}^F)), c_h) \\ &= \frac{\phi_{h_h}(x_i | B(T_I(I_{seg}^F))) + \phi_{h_h}(x_j | B(T_I(I_{seg}^F)))}{2}, \end{aligned} \quad (20)$$

and

$$\phi_{h_h}(x_i | B(T_I(I_{seg}^F)), c_h) = \frac{\sum_{\varepsilon \in B(T_I(I_{seg}^F))} w_\varepsilon(i) c_h(\varepsilon) d(i, \varepsilon)}{\rho_{B(T_I(I_{seg}^F))} \sum_{\varepsilon \in B(T_I(I_{seg}^F))} w_\varepsilon(i)}, \quad (21)$$

with w and ρ computed in the same way as for the patch boundary potentials with $B(T_I(I_{seg}^F))$ instead of B_m .

2. Adaptive Parameter Determination

An important aspect in energy minimization based segmentation methods is the determination of the parameters, most importantly, the value of weights λ_{ϕ_l} and λ_b applied to the unary and pairwise energy terms. Previously, these parameters were defined as constants and determined empirically [5,6]. In our method we apply simple assumptions to adaptively adjust this ratio based on statistics of I_{seg}^F and I_{seg}^B to ensure segmentation results are consistent with that given by the user.

The first assumption is that the more the pixel distributions H_{seg}^F and H_{seg}^B are separated, so will the distributions of H_I^F and H_I^B . Also, the more H^F and H^B are separated, the more they become a stronger criterion to distinguish between I^F and I^B . Therefore, we increase λ_{ϕ_l} according to the distance between H_I^F and H_I^B by

$$\lambda_{\phi_l} = -\log(D_l(H_I^F, H_I^B)). \quad (22)$$

The second is that if there are large gradients other than the gradients at the boundary, gradients do not provide a good criteria in segmenting the image. Therefore, we decrease λ_b according to the ratio between the average gradients at $B(I_{seg}^F)$ and the average gradients at edges of I_{seg} extracted by an edge detector.

3. Energy Minimization

As mentioned, the values of $\mathbb{X}$ are determined through minimization of the energy E of (1) based on the computed potentials. This is done by constructing a graph with edge capacities corresponding to the potentials and finding its min-cut following [5] and [15]. The additional potentials of the superpixels defined in (15) are incorporated in the graph by constructing auxiliary nodes connecting them to nodes corresponding to their constituent pixels following the theory recently proposed in [20].

4. Iterative Segmentation

While the segmentation process is iterated throughout the whole set of consistent scene images, the low-level and mid-level cues Q_l and Q_m are accumulated in order to enrich the information of both the foreground and background. Q_l is accumulated by averaging $\{H_{seg}^F, H_{seg}^B\}$, $\{H_{l,m}^F, H_{l,m}^B\}$, and $\{H_{l,h}^F, H_{l,h}^B\}$ by the weights c_l , c_m , and c_h . Therefore, the low-level cue $Q_l^{\neq xt}$ to be propagated to $I_{\neq xt}$ is defined as

$$Q_l^{\neq xt} = \{H_I^F, H_I^B\}, \quad (23)$$

where

그림 4. 고정된 물체를 움직이면서 촬영한 30개의 영상으로 이루어진 DINO 집합의 원 영상 (윗 줄) 및 영역화된 전경 (아래 줄)

Fig. 4. Samples of original images and segmented foreground for the DINO image set comprising of 16 images of a moving person taken at changing viewpoints

$$H_I(p) = \frac{c_l H_{I_{seg}}(p) + c_m H_{I_m}(p) + c_h H_{I_h}(p)}{c_l + c_m + c_h}. \quad (24)$$

Accumulation of mid-level cues is done by including relevant patches ω_I extracted from I and excluding irrelevant patches ω from the dictionaries Ω^F and Ω^B through the resampling process presented in [1]. Here, we omitted the subscript for Ω representing that the patch dictionaries do not depend only on I_{seg} and evolve in the iteration process depending on all images in the set.

V. Experimental Results

We present experimental results on a variety of consistent scene images, mainly focusing on images sets of an object or images of a person in motion. The images of the BUS set of Fig. 1 were collected from the VOC2006 database of the Pascal object recognition challenge [16]. Although there is a considerable amount of change in appearance for both foreground and background, the pro-

posed method produce consistent results. Fig. 4 presents another example of images of an object that may be used as the input for multi-view stereo, and its segmented foreground. We can see that due to abrupt changes in the appearance, errors may be propagated to adjacent images, specifically, at in the tail of the dinosaur of image #5, and therefore the segmentation becomes erroneous. But, this error may be overcome, as shown in image #7, by the confidence measures and compensation from other cues which together reinforce robustness.

Fig. 5 presents the results for images of a person in motion with varying viewpoints, where image set is comprised of 16 images. The proposed method produces accurate results except for image #11 where parts of the arm were misclassified as background. This is due to the fact that the arm was not visible in the image initially segmented by the user.

We provide a comparison between results obtained from the proposed method and results based only on mid-level cues presented in [1] on the KIM set of [25] in Fig. 6. We extract 20 frames from the video of total 185 frames, there

그림 5. 움직이는 사람을 촬영한 15개의 영상으로 이루어진 JE 집합의 원 영상 (윗 줄) 및 영역화된 전경 (아래 줄)

Fig. 5. Samples of original images and segmented foreground for the JE image set comprising of 16 images of a moving person taken at changing viewpoints

그림 6. [25]에 제안된 비디오로부터 추출된 KIM영상 집합에 대해 (a) 중간 레벨 큐만을 이용한 [1]의 기법과 (b) 제안하는 기법의 결과 비교. 제안하는 기법의 영역 경계가 더 정확함을 확인할 수 있다

Fig. 6. Comparison between results (a) based only on mid-level cues presented in [1] and results obtained by the proposed method for the KIM image set extracted from video of [25] (image numbers are assigned as those of [1]). We can see that the proposed method produces a more clean boundary

by decreasing the correlation between the frames and creating an image set similar to one would get from continuous shooting of a digital camera. We can see that by incorporating multiple cues helps to segment the image with a more

accurate boundary.

We also present quantitative analysis of the segmentation results obtained by the proposed method for 7 image sets, where 3 sets, DINO, BOY, and LADY, are of an object,

and 4 sets, JE, JE2, JE3, and KIM, are of a moving person. All image sets are comprised of images taken at varying viewpoints where each set is comprised of 30 (DINO), 27 (BOY), 24 (LADY), 15 (JE), 20 (JE2), 13 (JE3), and 20 (KIM) images. We evaluate the segmentation results against ground truth data obtained by manual segmentation for each image set. The segmentation rate is defined as the ratio of pixels that are correctly classified as foreground against the size of the foreground of the ground truth, while false positives are defined as the rate of pixels that are classified as foreground in the results while classified as background in the ground truth against the size of the image.

As we can see from table 1, the results of the BOY set and the JE2 set have the lowest segmentation rate and highest false positive set. This is mainly due to the error accumulation that occurs for these sets. The error accumulation is attributed to low gradients in the boundaries between the foreground and background as well as the fact that the foreground and background are comprised of similar colors, creating camouflage. This problem is difficult for any segmentation method, and despite these obstacles, our method is able to produce fairly accurate segmentation results. As for the other data sets, the proposed method produces highly accurate results.

표 1. 제안하는 기법의 영상 집합 별 영역화율과 오검출율

Table 1. Segmentation rate and false positive rate of proposed method averaged over image sets

Data Set	Segmentation Rate (%)	False Positive (%)
BOY	93.6	2.00
DINO	98.6	0.234
LADY	97.2	0.359
JE	96.0	0.108
JE2	91.7	2.96
JE3	99.3	0.315
KIM	99.6	0.234

For the results presented here, mean shift segmentation

[2] was applied to obtain superpixels, the Bhattacharyya distance was used for D_l , and the Earth Mover's Distance [22] was user for D_m . Also, the parameters other than λ_{ϕ_l} and λ_b are empirically determined as $\lambda_m = 1$ and $\lambda_{\phi_m} = \lambda_{\phi_h} = 3$ for all the results presented in this section. We note that this demonstrates the robustness of the proposed method to parameters, which is based on the adaptive selection process for the other parameters. We also note that the processes of extracting multiple cues and measuring their confidence increases computational cost, taking close to one minute for a 640×480 image. But, since the extracted cues may be extracted in parallel, we believe that this may be improved by utilizing multiple threads.

References

- [1] I. Lu, L., Hager, G.: A nonparametric treatment for location/segmentation based visual tracking. In: CVPR07. (2007)
- [2] Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. PAMI 24 (2002) 603 - 619
- [3] Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. IJCV 59 (2004) 167 - 181
- [4] Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI 22 (2000) 888 - 905
- [5] Boykov, Y., Funka Lea, G.: Graph cuts and efficient n-d image segmentation. IJCV 70 (2006) 109 - 131
- [6] Rother, C., Kolmogorov, V., Blake, A.: "grabcut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23 (2004) 309 - 14
- [7] Kumar, M., Torr, P., Zisserman, A.: Obj cut. In: CVPR05. (2005)
- [8] Freedman, D., Zhang, T.: Interactive graph cut based segmentation with shape priors. In: CVPR05. (2005) I: 755 - 62
- [9] Li, Y., Sun, J., Shum, H.Y.: Video object cut and paste. ACM Trans. Graph. 24 (2005)
- [10] Wang, J., Bhat, P., Colburn, R.A., Agrawala, M., Cohen, M.F.: Interactive video cutout. ACM Trans. Graph. 24 (2005) 585 - 94
- [11] Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., Rother, C.: Probabilistic fusion of stereo with color and contrast for bi-layer segmentation. IJCV 76 (2008)
- [12] Sun, J., Zhang, W., Tang, X., Shum, H.: Background cut. In: ECCV06. (2006) II: 628 - 41
- [13] Criminisi, A., Cross, G., Blake, A., Kolmogorov, V.: Bilayer segmentation of live video. In: CVPR06. (2006)
- [14] Yin, P., Criminisi, A., Winn, J., Essa, I.: Tree-based classifiers for

- bilayer video segmentation. In: CVPR07. (2007)
- [15] Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. PAMI 26 (2004) 1124 - 137
- [16] Homepage, T.P.V.O.C.: (<http://www.pascal-network.org/challenges/voc/>)
- [17] Rother, C., Minka, T., Blake, A., Kolmogorov, V.: Cosegmentation of image pairs by histogram matching: Incorporating a global constraint into mrfs. In: CVPR06. (2006) I: 993 - 000
- [18] Campbell, N., Vogiatzis, G., Hernandez, C., Cipolla, R.: Automatic 3d object segmentation in multiple views using volumetric graph-cuts. In: BMVC07. (2007)
- [19] Zheng, S., Tu, Z., Yuille, A.: Detecting object boundaries using low-, mid-, and high-level information. In: CVPR07. (2007)
- [20] Kohli, P., Kumar, M., Torr, P.: P3 & beyond: Solving energies with higher order cliques. In: CVPR07. (2007)
- [21] Li, S.Z.: Markov Random Field Modeling in Computer Vision. Springer-Verlag (1995)
- [22] Rubner, Y., Tomasi, C., Guibas, L.: The earth mover's distance as a metric for image retrieval. IJCV 40 (2000) 99 - 21
- [23] Kwon, D., Yun, I., Lee, S.: Efficient feature-based nonrigid registration of multi-phase liver ct volumes. BMVC (2008)
- [24] Xu, C., Prince, J.: Snakes, shapes, and gradient vector flow. T-IP 7 (1998) 359 - 69
- [25] Chuang, Y.Y., Agarwala, A., Curless, B., Salesin, D.H., Szeliski, R.: Video matting of complex scenes. ACM Transactions on Graphics 21 (2002) 243 - 48 Sepcial Issue of the SIGGRAPH 2002 Proceedings.

저 자 소 개

이 수 찬

- 2004년 : 서울대학교 전기공학부 학사
- 2004년 ~ 현재 : 서울대학교 전기공학부 박사과정 재학중
- 주관심분야 : 컴퓨터 비전, 3차원 모델링, 영상 영역화

윤 일 동

- 1989년 2월 : 서울대학교 제어계측공학과 공학사
- 1991년 2월 : 서울대학교 제어계측공학과 공학석사
- 1996년 8월 : 서울대학교 제어계측공학과 공학박사
- 1997년 3월 ~ 현재 : 한국외국어대학교 디지털정보공학과 교수
- 주관심분야 : 영상처리, 컴퓨터비전

이 상 옥

- 1973년 : 서울대학교 전기공학부 학사
- 1976년 : 미국 Iowa 주립대 전기공학과 석사
- 1980년 : 미국 Univ. of Southern California, 전기공학과 박사
- 1981년 : General Electric Co. 연구원
- 1983년 : M/A-COM Research Center 선임연구원
- 1983년 ~ 현재 : 서울대학교 전기공학부 정교수
- 주관심분야 : 영상처리, 컴퓨터 비전