

# SALIENCY-BASED VIDEO SEGMENTATION WITH GRAPH CUTS AND SEQUENTIALLY UPDATED PRIORS

Ken Fukuchi\*, Kouji Miyazato\*, Akisato Kimura†, Shigeru Takagi\* and Junji Yamato†

† NTT Communication Science Laboratories, NTT Corporation, Japan  
\* Department of Information and Communication Systems Engineering,  
Okinawa National College of Technology, Japan

## ABSTRACT

This paper proposes a new method for achieving precise video segmentation without any supervision or interaction. The main contributions of this report include 1) the introduction of fully automatic segmentation based on the maximum a posteriori (MAP) estimation of the Markov random field (MRF) with graph cuts and saliency-driven priors and 2) the updating of priors and feature likelihoods by integrating the previous segmentation results and the currently estimated saliency-based visual attention. Test results indicate that our new method precisely extracts probable regions from videos without any supervised interactions.

**Index Terms**— Video segmentation, saliency, Markov random fields, MAP estimation, graph cuts, Kalman filter.

## 1. INTRODUCTION

Extracting important (or meaningful) regions from videos is not only a challenging problem in computer vision research but also a crucial task in many applications including object recognition, video classification, annotation and retrieval. It can be formulated as a problem of binary segmentation, where important regions are considered “objects” and the remaining regions “backgrounds”. One of the most promising ways to achieve precise segmentation is the method proposed by Boykov et al. [1] called Interactive Graph Cuts. This method originated in the work of Greig et al. [2], where the exact maximum a posteriori (MAP) solution of a two label pairwise Markov random field (MRF) can be obtained by finding the minimum cut on the equivalent graph of the MRF. More recently, several approaches for extending it to video segmentation have been proposed. For example, Kohli and Torr [3] described an efficient algorithm for computing MAP estimates for dynamically changing MRF models, and tested it on the video segmentation problem.

Although the above approaches are promising, they all pose a critical problem in that they have to provide segmentation cues (seeds) manually and carefully. Such manual labeling is occasionally infeasible. The development of fully automatic segmentation methods has been strongly expected. The use of saliency-based human visual attention models is one of the most promising approaches in this respect. The first biologically plausible model for explaining the human attention system was proposed by Koch and Ullman [4], and late implemented by Itti et al. [5]. This model analyzes still images to produce primary visual features, which are combined to

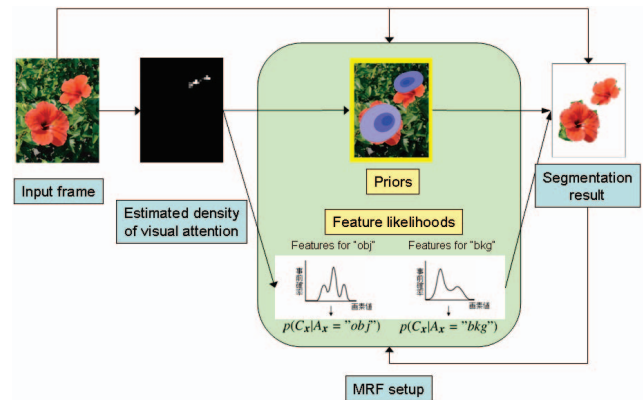


Fig. 1. Framework of proposed method

form a saliency map that represents the relevance of visual attention. Later, we have proposed a stochastic approach for estimating human visual attention [6, 7, 8] that tackled the fundamental problem of the previous attention models related to the non-deterministic properties of the human visual system. Such models would be helpful for automatically providing segmentation seeds.

We propose a novel approach for achieving video segmentation based on visual saliency. Our main contributions are as follows: 1) We introduce MAP-based frame-wise segmentation with graph cuts where priors for segmentation are provided based on visual saliency. This approach is closely related to the work undertaken by Fu et al. [9] for still image segmentation. 2) We develop a new technique for estimating and updating priors and feature likelihoods. We integrate the prior derived from the segmentation results for the previous frames and the prior derived from the saliency at the current frame by using a Kalman filter. The feature likelihood can be also estimated by combining two feature likelihoods, one obtained from the segmentation results for the previous frames and the other from the saliency calculation for the current frame.

## 2. FRAMEWORK

Fig. 1 depicts the framework of the proposed method for extracting salient regions from videos.

First, the visual attention density is calculated from each frame of an input video via a saliency-based human visual attention model. Although any kind of attention model can be employed, we utilize our attention model [6, 8] to compute the human visual attention

Contact address: ic041236@edu.okinawa-ct.ac.jp, akisato@ieee.org

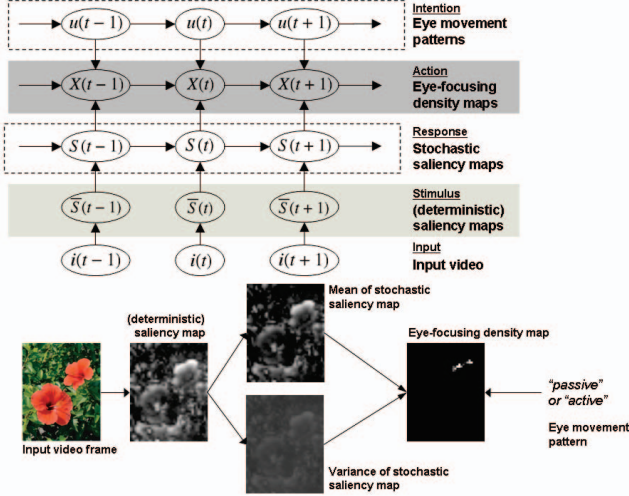


Fig. 2. Estimation of human visual attention through a stochastic attention model

density. Section 3 describes how to estimate human visual attention with the proposed method.

Next, a Markov random field (MRF) model for segmentation is prepared, where each hidden state corresponds to the label of a position representing an “object” or “background”, and an observation is a frame of the input frame. The density calculated in the previous step can be utilized for estimating the priors of objects/backgrounds and the feature likelihoods of the MRF. When calculating priors and likelihoods, the regions extracted from the previous frames are also available. Section 5 focuses particularly on how to determine and update priors and feature likelihoods based on the density of visual attention and previous segmentation results.

Once the MRF is constructed, salient regions can be obtained as the MAP solution of the MRF. When estimating the MAP solution, segmentation methods based on graph cuts [1] can be employed. Section 4 presents the segmentation method for frame segmentation.

### 3. ESTIMATION OF HUMAN VISUAL ATTENTION

Fig. 2 shows the framework for estimating human visual attention. We used our stochastic attention model [6, 8].

First, a saliency map is calculated from each frame of the input video with the method proposed by Itti et al. [5]. Our implementation utilized intensity, color opponents, orientation and motion information as fundamental features. Then, a stochastic representation of the saliency map is computed through a Kalman filter, where the saliency map is utilized as the observation of the filter. We call this stochastic representation as a stochastic saliency map. Each pixel of the stochastic saliency map is expressed by a Gaussian density. The density of human visual attention can be directly calculated from the stochastic saliency map by introducing the principle of the signal detection theory [10], namely, the position at which stochastic saliency takes its maximum value is the eye focusing position. Since each pixel of the stochastic saliency map is expressed by a Gaussian, we can calculate the visual attention density for each pixel such that the saliency value has its maximum value at that pixel. The model also incorporates another property, namely that eye movements may be affected by a cognitive state. The cognitive state is represented as

an eye movement pattern in this model. Two typical eye movement patterns, passive and active, are found when a person is watching a video. By introducing the eye movement patterns, eye movements can be modeled with a hidden Markov model. Finally, by integrating the density coming from the stochastic saliency map and the eye movement pattern, we can obtain the final density of visual attention, which is called the eye focusing density map (EFDM).

### 4. IMAGE SEGMENTATION BY GRAPH CUTS

This section describes the supervised image segmentation technique based on graph cuts proposed by Boykov et al. [1].

We start by describing MRFs for image segmentation. Consider a set of random variables  $\mathbf{A} = \{A_x\}_{x \in I}$  defined on a set  $I$  of coordinates. Each random variable  $A_x$  takes a value  $a_x$  from the set  $\mathcal{L} = \{0, 1\}$ , which corresponds to a background (0) and an object (1), respectively. A MAP-based MRF estimation can be formulated as an energy minimization problem where the energy corresponding to the configuration  $\mathbf{a}$  is the negative log likelihood of the joint posterior density of the MRF,  $E(\mathbf{a}|D) = -\log p(\mathbf{A} = \mathbf{a}|D)$ , where  $D$  represents the input image. The energy function consists of likelihood and prior terms defined as follows:

$$E(\mathbf{A}|D) = \sum_{x \in I} \left\{ \psi_1(D|A_x) + \xi_1(A_x) + \sum_{y \in N_x} (\psi_2(D|A_x, A_y) + \xi_2(A_x, A_y)) \right\}, \quad (1)$$

where  $N_x$  is a neighborhood system to the position  $\mathbf{x}$ ,  $\psi_i(D|\cdot)$  ( $i = 1, 2$ ) is a likelihood term and  $\xi_i(\cdot)$  is a prior term. The first likelihood term  $\psi_1(D|A_x)$  is the negative log likelihood which imposes individual penalties for assigning label  $l \in \mathcal{L}$  to pixel  $\mathbf{x}$ . It is given by  $\psi_1(D|A_x) = -\log p(C_x|A_x)$ , where  $C_x$  is the RGB value at the position  $\mathbf{x}$ . The likelihood  $p(C_x|A_x)$  of the RGB values can be modeled as a Gaussian mixture model (GMM), and estimated with a standard EM algorithm. The first prior term  $\xi_1(A_x)$  represents how the position is likely to an object, and it is given by  $\xi_1(A_x) = -\log p(A_x)$ . The prior density  $p(A_x)$  can be determined by labels manually given from users as

$$p(A_x = 1) = \begin{cases} 1 & \text{The label "obj" is given at } \mathbf{x} \\ 0 & \text{The label "bkg" is given at } \mathbf{x} \\ 0.5 & \text{No label provided at } \mathbf{x} \end{cases}$$

$$p(A_x = 0) = 1 - p(A_x = 1).$$

The second prior term  $\xi_2(A_x, A_y)$  takes the form of a generalized Potts model as  $\xi_2(A_x, A_y) = K$  only if  $A_x \neq A_y$ . The second likelihood term  $\psi_2(D|A_x, A_y)$  reduces the cost for two labels, which differs in proportion to the difference between the intensity values of their corresponding positions.

$$\psi_2(D|A_x, A_y) \propto -\exp \left\{ -\frac{(I_x - I_y)^2}{2\sigma^2} \right\} \cdot \frac{1}{\|\mathbf{x} - \mathbf{y}\|} \quad \text{if } A_x \neq A_y,$$

where  $I_x$  denotes the intensity at the pixel  $\mathbf{x}$ .

The MRF configuration  $\hat{\mathbf{a}}$  with the least energy corresponds to the MAP solution of the MRF. The minimization of energies can be performed by finding the minimum cut on an equivalent graph of

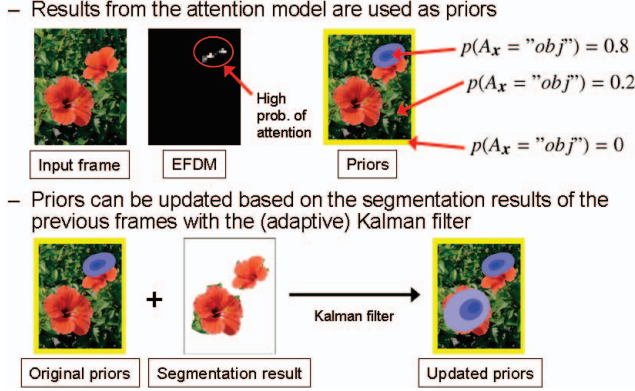


Fig. 3. The main contributions of our proposed method

the MRF. Each random variable  $A_x$  of the MRF is represented by a vertex  $v_x$  in this graph. The edges of a vertex  $v_x$  are connected to the vertexes in its neighborhood  $N_x$ . These edges are called as neighborhood links (n-links). The cost  $c(v_x, v_y)$  associated with the n-link  $(x, y)$  connecting vertexes  $v_x$  and  $v_y$  is given by the sum of the second prior and likelihood terms. The two labels “obj” and “bkg” are represented by special vertexes, namely the source  $s$  and the sink  $t$ . They are connected to all other vertexes by edges called terminal links (t-links). The costs  $c(s, v_x)$  and  $c(t, v_x)$  of t-links are given by the sum of the first prior and likelihood terms. A graph cut in this graph, which separates the source and the sink, defines the MAP configuration  $\hat{a}$  of the MRF.

## 5. MAIN CONTRIBUTIONS

### 5.1. Attention-based priors and likelihoods

Fig. 3 shows a sketch of the contributions. First, our method provides a way to calculate the prior and likelihood terms of the energy function (Eq. 1) without any manually provided labels. Instead, we utilize the density of visual attention calculated by the procedure shown in Section 3.

The first prior term  $\xi_1(A_x = 1)$  is the negative log of the prior density obtained from the EFDM (cf. Section 3). The EFDM is modeled by a GMM, and the model parameter is estimated with the EM algorithm. The estimated GMM density represents the first prior density  $p(A_x = 1)$ . Exceptionally, the prior density on the edge of each frame is assumed to be  $p(A_x = 1) = 0$  since some of the background regions are expected to be at the frame edge. The top of Fig. 3 shows an example of priors.

The first likelihood term  $\psi_1(D|A_x)$  is the negative log likelihood of the RGB values obtained in a similar way as the Interactive Graph Cuts [1]. In the Interactive Graph Cuts, samples are selected from the manually-labeled pixels. In contrast, our proposed method utilizes all the pixels for estimating the likelihood  $\psi_1(D|A_x)$ , where samples are weighted by the first prior density  $p(A_x = 1)$ .

### 5.2. Updating priors and likelihoods

The second contribution provided by our method is that it offers a way to update the prior and likelihood terms according to the segmentation results derived from the previous frames and the density

of visual attention calculated from the current frame. Here, we introduce a notation  $\mathbf{A}_t = \{A_{x,t}\}_{x \in I}$  ( $t = 0, 1, \dots$ ) for representing the MRF configuration at time  $t$ .

To update the first prior density  $p(A_{x,t} = 1; t)$  at time  $t$ , we introduce the idea of the Kalman filter, where the prior density  $q(A_{x,t}; t)$  derived from solely the EFDM at time  $t$  is considered to be the observation at time  $t$ . We assume the following two relationships:

$$\begin{aligned} p(p(A_{x,t} = 1; t)) &= \mathcal{G}(p(A_{x,t} = 1; t); f(\hat{a}_{x,t-1}), \sigma_1), \\ p(q(A_{x,t} = 1; t)) &= \mathcal{G}(q(A_{x,t} = 1; t); p(A_{x,t} = 1; t), \sigma_2), \end{aligned}$$

where  $\hat{a}_{x,t}$  is the estimated label on position  $x$  at time  $t$ ,  $\sigma_i$  ( $i = 1, 2$ ) is a model parameter given in advance,  $\mathcal{G}(s; \bar{s}, \sigma)$  is the Gaussian density with argument  $s$ , mean  $\bar{s}$  and variance  $\sigma^2$ , and  $f(a_x)$  is a deterministic function that converts a label to a real value. These equations imply that the prior term at time  $t$  depends on both the density of visual attention at time  $t$  and the segmentation result at time  $t-1$ . The ML estimate  $\hat{p}(A_{x,t} = 1; t)$  of the first prior density at time  $t$  can be obtained by the Kalman filter as follows:

$$\begin{aligned} \hat{p}(A_{x,t} = 1; t) &= \frac{\sigma_{\xi_1}^2(t)}{\sigma_2^2 + \sigma_{\xi_1}^2(t-1)} f(\hat{a}_{x,t-1}) \\ &\quad + \frac{\sigma_{\xi_1}^2(t)}{\sigma_1^2} q(A_{x,t} = 1; t), \\ \sigma_{\xi_1}^2(t) &= \frac{\sigma_1^2 \cdot (\sigma_2^2 + \sigma_{\xi_1}^2(t-1))}{\sigma_1^2 + \sigma_2^2 + \sigma_{\xi_1}^2(t-1)}, \end{aligned}$$

The ML estimate  $\hat{\xi}_1(A_{x,t} = 1; t) = -\log \hat{p}(A_{x,t} = 1; t)$  derived from the above procedure is used as a new prior term.

The first likelihood term  $\psi_1(D|A_{x,t}; t)$  at time  $t$  can also be derived from the previous segmentation result and the current density of visual attention. As shown in Section 4, this term is the negative log likelihood  $-\log p(C_{x,t}|A_{x,t})$  of the RGB values for a given label. The likelihood  $p(C_{x,t}|A_{x,t}; t)$  can be modeled with a mixture of two GMMs.

$$\begin{aligned} p(C_{x,t}|A_{x,t}; t) &= \lambda q_1(C_{x,t-1}|A_{x,t-1}) + (1-\lambda) q_2(C_{x,t}|A_{x,t}), \end{aligned}$$

where  $\lambda$  is the mixture ratio given in advance,  $q_1(C_{x,t-1}|A_{x,t-1})$  is estimated from samples selected from the “object” region of the segmentation result at time  $t-1$ , and  $q_2(C_{x,t}|A_{x,t})$  is estimated from samples weighted by the first prior density at time  $t$ .

## 6. TEST RESULTS

To verify the effectiveness of the proposed method, we conducted video segmentation for several video clips. We compared our new method with a frame-wise segmentation algorithm based on the Interactive Graph Cuts, which is equivalent to the new method without sequentially updated priors.

Fig. 4 shows several segmentation examples. Owing to the limited spaces, detailed results have been placed on the web<sup>1</sup>, including input videos, segmentation results (namely videos after segmentation), videos of visual attention density, and prior videos. The results are qualitatively good and largely agree with perceptual boundaries.

<sup>1</sup> <http://www.brl.ntt.co.jp/people/akisato/saliency3.html>

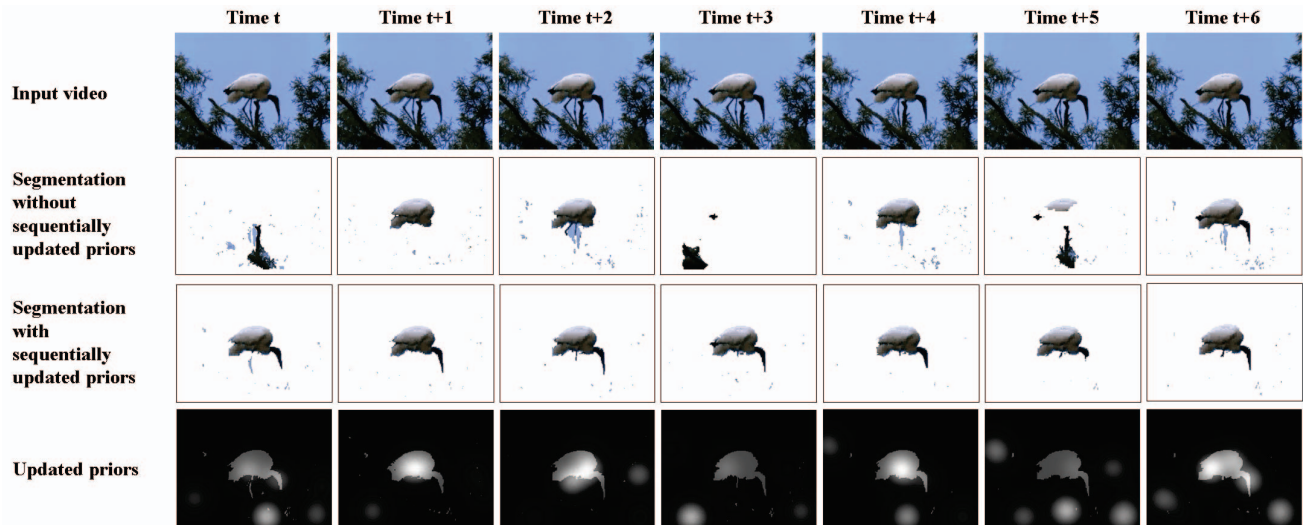


Fig. 5. Comparison of the proposed method with and without sequentially updated priors. From the top row to the bottom row: Input video, segmentation result without and with sequentially updated priors, the first prior term derived by the Kalman filter.

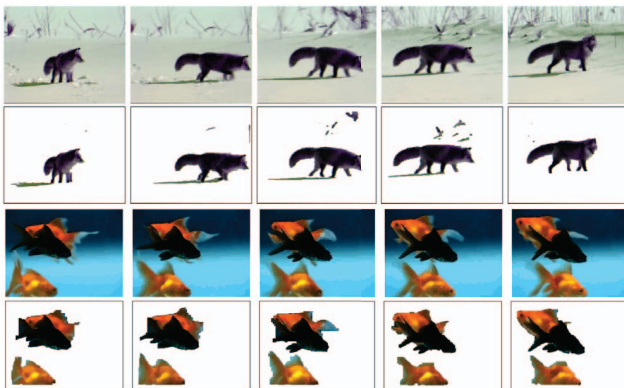


Fig. 4. Segmentation results.

Fig. 5 compares the new method with and without sequentially updated priors shown in Section 5.2. As shown in the second row of Fig. 5, the segmented regions were randomly switched as a result of the shifts of attention without sequentially updated priors. This problem could be appropriately eliminated with the use of sequentially updated priors as shown in the third row of Fig. 5.

## 7. CONCLUSION

This report have proposed a new method for achieving precise video segmentation without any supervised interactions. The main contributions included 1) the introduction of MAP-based frame-wise segmentation with graph cuts and saliency-driven priors, and 2) the technique for updating priors and likelihoods with a Kalman filter. The results showed that our algorithm extracted probable regions appropriately. Future work will include efficient computation of graph cuts by exploiting the fact that the change in the MRF is relatively small from one time instant to another (cf. Kohli and Torr [3]), quantitative analysis of the proposed method, and a unified statistical framework for attention estimation and salient region extraction.

## 8. REFERENCES

- [1] Y. Boykov and G.F. Lea, “Graph cuts and efficient N-D image segmentation,” *IJCV*, vol. 70, no. 2, pp. 109–131, 2006.
- [2] D. Greig, B. Porteous, and A. Seheuit, “Exact maximum a posteriori estimation for binary images,” *Royalstat*, vol. B:51, no. 2, pp. 271–279, 1989.
- [3] P. Kohli and P. Torr, “Dynamic graph cuts for efficient inference in markov random fields,” *IEEE Trans. PAMI*, vol. 29, no. 12, pp. 2079–2088, 2007.
- [4] C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the underlying neural circuitry,” *Human Neurobiology*, vol. 4, pp. 219–227, 1985.
- [5] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. PAMI*, vol. 20, no. 11, pp. 1254–1259, November 1998.
- [6] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino, “A stochastic model of selective visual attention with a dynamic Bayesian network,” in *Proc. ICME2008*, June 2008, pp. 1076–1079.
- [7] A. Kimura, D. Pang, T. Takeuchi, J. Yamato, and K. Kashino, “Dynamic Markov random field for stochastic modeling of visual attention,” in *Proc. ICPR2008*, December 2008, p. Mo.BT8.35.
- [8] K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, “Real-time estimation of human visual attention with mcmc-based particle filter,” to appear in *Proc. ICME2009*, June 2009.
- [9] Y. Fu, J. Cheng, Z. Li, and H. Lu, “Saliency cuts: An automatic approach to object segmentation,” in *Proc. ICPR2008*, 2008.
- [10] M. P. Eckstein, J. P. Thomas, J. Palmer, and S. S. Shimozaki, “A signal detection model predicts effects of set size on visual search accuracy for feature, conjunction, triple conjunction and disjunction displays,” *Perception and Psychophysics*, vol. 62, pp. 425–451, 2000.