THE INTERNATIONAL JOURNAL OF
SCIENCE & TECHNOLEDGE

Saliency Based Object Extraction in Single Concept Video Frames

Priya Devi S.
Student, Department of Computer Science and Engineering
KSR College of Engineering, Tamil Nadu, India
Iniya K.
Student, Department of Computer Science and Engineering
KSR College of Engineering, Tamil Nadu, India
Surendheran A. R.
Assistant Professor, Department of Computer Science and Engineering
KSR College of Engineering, Tamil Nadu, India

Abstract:
This paper presents saliency scheme based object extraction method to extract and fragment the objects from frames. At first visual and motion saliency feature information map are derived from input frames, and then that feature cosaliency informations are integrated with their color models. Then conditional random field (CRF) was applied for automatic object extraction. It will able to produces satisfactory result.

Key words*: Object extraction, Visual and Motion saliency, Cosaliency, Conditional random field (CRF)*

1. Introduction

Extracting objects from video frames has been important task in lot of applications such as video annotation, event recognition, and object retrieval. It will different from other object extraction methods. It will able to handle in dynamic, unpredicted clutter backgrounds. Mostly object extraction will do by supervised or unsupervised ways.

While [1, 2] requires user interaction, [3, 4] requires training datasets on the objects of interest. This method unsupervised method based. So, it does not require any user interaction and the training datasets for extraction. At first Itti et al [5] proposed visual saliency based on the visual attention system information are captured from various applications. For example spectral residual of FFT are computed in [6], in [7] saliency object models are determined based on multi-scale contrast and color information. [8] Utilize averaged color and information from neighbor pixels.

This paper will extend the concept of image based visual and motion saliency. Visual and motion saliency feature information are extracted from input video frames, that cosaliency feature information are integrated with their associated color models. And then CRF will be applied for automatic object extraction. This method does not require any user interaction and training datasets. And it will able to deal with multiple object instance not only single object instance. Figure 1 shows the outline of our work.

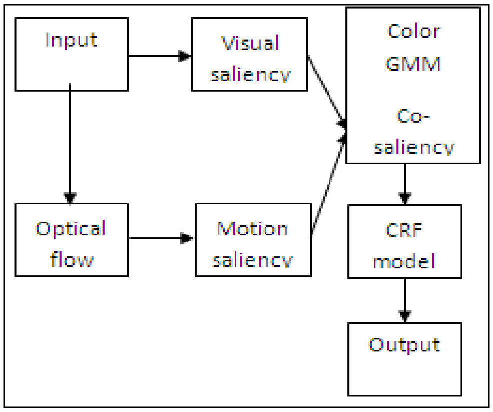

Figure 1: Outline of our framework

2. Saliency Significant Video Object Extraction

In this paper to integrate both visual and motion saliency information transversely video frames, and the extracted cosalient regions with the consequent color models can be applied to recognize the foreground objects of interest.

2.1. Determination of Visual Saliency

Stimulated by [7], this paper advance color and contrast information in each video frame for detecting visual saliency. In our work, we apply Turbopixels future by [8] to fragment each video frame, and we perform saliency recognition at the superpixel. Compare to [7], this procedure results in edge-preserving superpixels with parallel size, and the use of these superpixels provide better visual saliency consequences. To determine the visual saliency of the extracted superpixels, we equivalently quantize every of their RGB color channels into 12 dissimilar bins, and further convert them into CIELAB color space due to its effectiveness in instead of color contrast information, as suggested by [4, 7]. First concern the saliency detection algorithm of [7], and evaluate the initial visual saliency of a superpixel r_k as follows:

$$\boxed{\begin{aligned} S(r_k) &= \sum_{r_k \neq r_i} \exp(D_s(r_k, r_i)/\sigma_s^2)\omega(r_i) D_r(r_k, r_i) \\ &\approx \sum \exp(D_s(r_k, r_i)/\sigma_s^2) D_r(r_k, r_i) \end{aligned}} \quad (1)$$

where D_s is the Euclidean distance among the centroid of r_k and that of its neighboring superpixels r_i, σ_s control the width of space kernel, and $\omega(r_i)$ is the mass of the neighbor superpixel r_i, which is relative to the amount of pixels in r_i. The last term $D_r(r_k,r_i)$ measures the color variation among r_k and r_i (also in terms of Euclidean distance). Image segment which are nearer to the most salient region should be further visually significant than farther away ones. As a result, behind calculate the saliency score $S(r_k)$ for each patch r_k, additional update the saliency S(i) for each pixel i by:

$$\boxed{\hat{S}(i) = S(i) * (1 - \text{dist}(i)/\text{dist}_{\text{max}})} \quad (2)$$

where S(i) is innovative saliency score resultant by (1) and dist(i) measures the Euclidian distance between the pixel I and its adjacent salient point. We note that dist_{max} in (2) is a constant representing the maximum expanse from a pixel of interest to its nearest salient point within an image. As illustrated in Figure 2(b), our approach can smooth the resultant visual saliency and improve possible nearby salient regions.

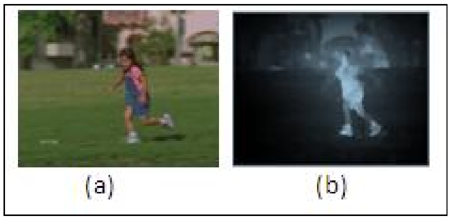

Figure 2. (a)Original video frame. (b) visual saliency map

2.2 Determination of Motion Saliency

In order to utilize the sequential characteristics of the input video, we detain the motion in sequence by scheming optical flow. To perform more precise optical flow assessment, we apply dense optical-flow with both forward and backward propagation [16] at every frame. Thus, each moving pixel i at frame t is determined by:

$$\boxed{m_t(i) = (\hat{m}_{t-1,t}(i) + \hat{m}_{t,t+1}(i))/2,} \quad (3)$$

Once such motion in sequence is extracted, we establish the motion saliency across video frames to recognize the prospective foreground objects of significance. This is achieved by manipulative the saliency of the above optical flow using (1), and the concluding motion saliency map can be formed (one for each frame).

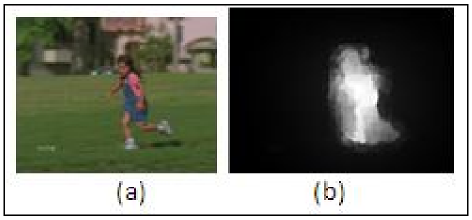

Figure 3: (a)Original video frame. (b) Motion saliency map

2.3 Derivation of a Cosaliency Map for VOE

Once both visual and motion saliency maps are resultant from an input video, we merge both in sequence and compute the connected cosaliency for later VOE purposes, as described in Figure1. It is importance noting that, when coiffure both saliency results, it is preferable to conserve the limit information since our vital goal is to fragment the objects of attention (not a region or a mask). Consequently, a edge-preserving term is introduced into our cosaliency determination:

$$\begin{aligned}&\text{Cosaliency A} = (Motion + Visual) * (EdgePreserving)\\ &= \{(M * vs/(ms + vs)) + (\hat{S} * ms/(ms + vs))\} * \exp(\hat{S}).\end{aligned} \quad (4)$$

In the above equation, A is the cosaliency map, ms and vs point out the sums of the motion and visual saliency values, correspondingly. We note that (4) leverages both saliency maps while the power for each is inversely proportional to the saliency values of the connected pixels. These improve the trouble of consequential in a biased cosaliency map if one of the saliency information is more dominant. As discussed above, the last exponential term on visual saliency is to preserve the visual boundary of the candidate foreground objects. Figure 4 shows several cosaliency detection examples using our proposed method, including comparisons with other state-of the- art image-based visual saliency detection approaches.

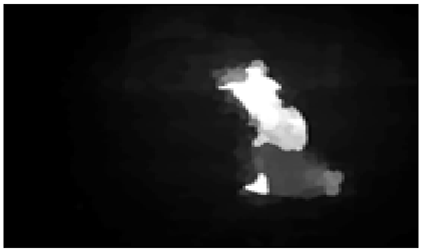

Figure 4: Cosaliency map

2.4 Learning of Color Cues Across Video Frames

While our cosaliency recognition process identify visual and motion salient regions in a video sequence, the saliency information itself is not sufficient for distinguishing foreground from background regions across videos. Therefore, based on the cosaliency detection results, we further consider saliency induced color information and construct both foreground and background color models for VOE purposes.

To make the above color models, we obtain the cosaliency map of every video frame and consider the pixels contained by as saliency-induced foreground and those in the enduring parts as background. Similar to [7], we apply the Gaussian Mixture Model (GMM) with ten Gaussian components for everyone to model the connected CIELAB color circulation. For later feature fusion purposes, we use a single energy term E^C to represent both foreground (E^{CF}) and background (E^{CB}) color models, and E^C is defined as:

$$E^{C} = E^{CF} - E^{CB} \quad (5)$$

In our proposed framework, we renew both color models for every frame, so with the purpose of these color models will be more adaptive to disparity of foreground or background diagonally video frames. It is also worth noting that, a prior VOE work in [5] only focused on modeling the foreground color information.

3. Conditional Random Field

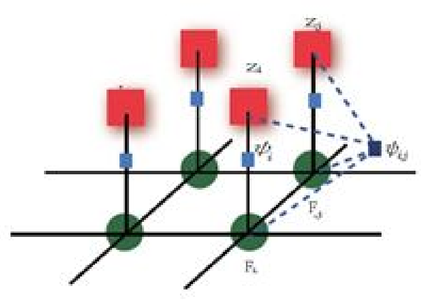

Figure 5: CRF for Object Segmentation

Utilizing an undirected graph, conditional random field (CRF) [9] is an authoritative method to approximate the structural information of a set of variables with the related interpretation. For video foreground object segmentation, CRF has been applied to calculate the label of each experimental pixel in an image *I* [5], [6]. As illustrated in Fig. 5, pixel *i* in a video frame is related with remark z_i, while the unknown node F_i point out its related label. In this framework, the label F_i is calculated by the observation z_i , while the spatial coherence between this output and neighboring remarks z_j and labels F_j are concurrently taken into consideration. Therefore, predicting the label of an observation node is equivalent to exploit the following posterior probability function

$$\boxed{p(F|I,\psi) \propto \exp\left\{-\left(\sum_{i\in I}(\psi_i) + \sum_{i\in I, j\in\text{Neighbor}}(\psi_{i,j})\right)\right\}} \quad (6)$$

where ψ_i is the unary term which infers the likelihood of F_i with observation z_i . $\psi_{i,j}$ is the pair wise term relating the correlation among neighboring pixels z_i and z_j, and that among their predicted output labels F_i and F_j . Note that the observation *z* can be represented by a particular feature or a grouping of various types of features. To solve a CRF optimization problem, one can convert the above problem into an energy minimization task, and the object energy function *E* of (6) can be derived as

$$\boxed{\begin{aligned} E &= -\log(p) \\ &= \sum_{i\in I}(\psi_i) + \sum_{\substack{i\in I \\ j\in\text{Neighbor}}}(\psi_{i,j}) \\ &= E_{\text{unary}} + E_{\text{pairwise}}. \end{aligned}} \quad (7)$$

In our proposed framework, we define the shape energy function E^S in terms of shape likelihood $X_t{}^S$ (derived by (5)) as one of the unary terms

$$\boxed{E^{\mathcal{S}} = -w^s \log(\widehat{X}_t^{\mathcal{S}}).} \quad (8)$$

In addition to shape information, we need incorporate visual saliency and color cues into the introduced CRF framework. As discussed earlier, we derive foreground and background color models for object extraction, and thus the unary term E^C describing color information is defined as follows:

$$\boxed{E^{\mathcal{C}} = w^c(E^{\mathcal{CF}} - E^{\mathcal{CB}}).} \quad (9)$$

Note that the foreground and background color GMM models G^C and $G^C{}_b$ are utilized to originate the related energy terms E^{CF} and E^{CB}, which are calculated as

$$\boxed{\begin{cases} E^{\mathcal{CF}} = -\log\left(\sum_{i\in I} G^{C_f}(i)\right) \\ E^{\mathcal{CB}} = -\log\left(\sum_{i\in I} G^{C_b}(i)\right). \end{cases}}$$

As for the visual saliency cue at frame *t*, we convert the visual saliency score derived in (2) into the following energy term E^V :

$$\boxed{E^{\mathcal{V}} = -w^v \log(\widehat{S}_t).} \quad (10)$$

We note that in the above equations, parameters w^s , w^c, and w^v are the weights for shape, color, and visual saliency cues, correspondingly. These weights organize the contributions of the related energy terms of the CRF model for performing VOE. It is also worth noting that, Liu and Gleicher [10] only considers the construction of foreground color models for VOE. As verified by [11], it can be concluded that the disregard of background color models would limit the performance of object extraction, since the only use of foreground color model might not be enough for characteristic between foreground and background regions. In the proposed object extraction framework, we now utilize multiple types of visual and motion salient features for object extraction.

4. Conclusion

This paper presents a saliency inspired object extraction method to segment the foreground objects of interest from video frame. Different from prior supervised object extraction approaches and those requiring user interactions, here provided an unsupervised VOE framework which aims at extracting visual and motion saliency information from the input video. With the saliency induced foreground and background color models, a CRF is utilized to integrate all the extracted features and thus object extraction problems can be solved automatically. Our method not only allows both foreground and a background region with significant motion across video frames, one of the major advantages is the capability to extract multiple object categories and instances in a video.

5. References

1. L. Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE PAMI, 1998.
2. X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in IEEE CVPR, 2007.
3. T. Liu et al., "Learning to detect a salient object," in IEEE CVPR, 2007.
4. R. Achanta and S. Susstrunk, "Saliency Detection using Maximum Symmetric Surround," in IEEE ICIP, 2010.
5. S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in IEEE CVPR, 2010.
6. Z. Liu, Y. Xue, L. Shen, and Z. Zhang, "Nonparametric saliency detection using kernel density estimation," in IEEE ICIP, 2010.
7. M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in IEEE CVPR, 2011.
8. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, and S. J. Dickinson, "Turbopixels: Fast superpixels using geometric flows," IEEE PAMI, 2009.
9. D. Tsai, M. Flagg, and J. M. Rehg, "Motion coherent tracking with multi-label mrf optimization," BMVC, 2010.
10. F. Liu and M. Gleicher, "Learning color and locality cues for moving object detection and segmentation.," in IEEE CVPR, 2009.
11. K.-C. Lien and Y.-C. F. Wang, "Automatic object extraction in singleconcept videos," in IEEE ICME, 2011