A User-Attention Based Focus Detection Framework and Its Applications

Chia-Chiang Ho, Wen-Huang Cheng, Ting-Jian Pan, Ja-Ling Wu

Communication and Multimedia Laboratory,
Department of Computer Science and Information Engineering,
National Taiwan University,
No. 1, Roosvelt Rd. Sec. 4, Taipei, Taiwan

Abstract

In this paper, a generic user-attention based focus detection framework is developed to capture user focus points for video frames. The proposed framework considers both bottom-up and top-down attentions, and integrates both image-based and video-based visual features for saliency map computation. For efficiency purpose, the number of adopted features is kept as few as possible. The realized framework is extensible and flexible in integrating more features with a variety of fusion schemes. One application of the proposed framework, the user-assisted spatial resolution reduction, has also been addressed.

1. Introduction

Attention refers to the ability of one human to focus and concentrate upon some visual or auditory "object", by carefully watching or listening. Assuming limited processing resources of one human, attention also refers to the allocation of these resources. Here, the resource can refer to either neurological or cognitive resource. The former is often referred as bottom-up attention and the later top-down attention. We can roughly say that bottom-up attention models what people are attracted to see, and top-down attention models what people are willing to see. In short, these two models can be summarized briefly as follows:

Bottom-up attention: The bottom-up attention can be modeled as an integration of different measurable, low-level image features [1]. Koch and Ullman proposed the first neurally plausible computational architecture of bottom-up attention model in 1985 [2]. Later researches on bottom-up attention generally followed this pioneering architecture. Nowadays, the bottom-up attention model proposed by Itti et al. draws great attention [3].

Top-down attention: Although the bottom-up attention model may capture the deployment of attention within the first few hundreds of milliseconds after presenting the visual scene, a complete attention model must consider top-down, task-oriented influences as well. Based on bottom-up attention model, there are researches that integrated the concept of top-down attention for object recognition tasks [4][5][6]. For general purposes, top-down attention is usually modeled by detecting some meaningful (semantic) objects or video features. For example, in [7], top-down attention was modeled by face detection and camera motion detection.

In this paper, we integrate both bottom-up and top-down features in our user attention model. Then, we propose a general user-attention based framework to detect users' focuses in video frames. Our work may provide another point of view for solving some content-aware problems. We present one application and related experimental results of the proposed focus detection framework: the user-assisted spatial resolution reduction which aims at obtaining better spatial resolution reduction of the input video instead of direct sub-sampling. Modeling User attention may also benefit other applications, such as video encoding, surveillance, watermarking, and video summarization [7].

The rest of this paper is organized as follows: Section 2 introduces the proposed user-attention based focus detection framework. Section 3 discusses the user-assisted spatial resolution reduction. Section 4 reports some experimental results. Finally, Section 5 presents the conclusion and future works.

2. The Proposed User-Attention Based Focus Detection Framework

In this section, we propose a user-attention based focus detection framework, which combines both bottom-up and top-down attention features, such as intensity, color, motion and face. Without fully semantic understanding of video content, the proposed framework provides us another way to benefit many content-based applications. Meanwhile, the system is carefully designed to deal with speed issue which real-time applications concerned about. Fig. 1 shows the snapshot and illustrations of the implemented focus detection framework.

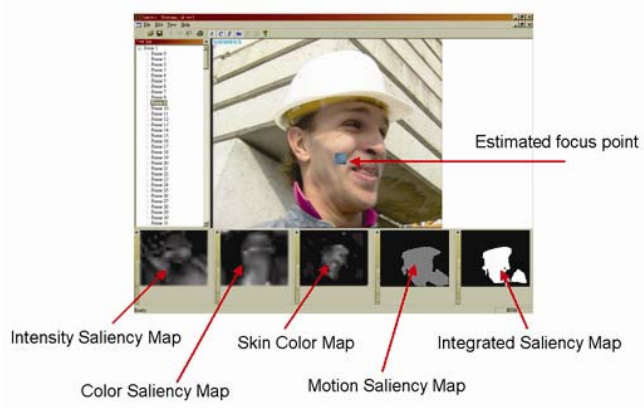

Figure 1. The operational snapshot of the proposed user-attention based focus detection framework.

2.1 Attentional Visual Features Calculation

We model computable attentional visual features into three levels, say, low level, medium level and high level.

Low-level features (intensity and color): Features belonging to this level correspond to the so-called early visual features in the biological vision, including but not limited to intensity contrast, color opponency, and orientation. Speed is a major concern in our system, so only intensity and color are used in our attention model. It is worthy of mentioning that in our observation, most of salient regions (or objects) found through computation of different early visual features might actually be similar. In our system, intensity and color features are calculated in the YCbCr color space, which is used by most of video coding standards.

The generation of adopted feature maps is similar to that of the Itti's method [8]. A map normalization operation is then applied to each feature map, which globally promotes maps in which a small number of strong peaks are presented and globally suppresses those maps with many peaks. Another effect of normalization operation is to let all feature maps share a common dynamic range. The normalization operation is performed by:

a) Finding out the maximal and minimum values, $MinVal$ and $MaxVal$, of the feature map, and then calculate a threshold value δ, which is defined as
$$MinVal + (MaxVal - MinVal)/10. \quad (1)$$

b) Counting the average value of pixels with values larger than δ, V_δ.

c) Calculating the map scaling factor as
$$F = \frac{MaxVal - V_\delta}{MaxVal - MinVal} \times \frac{255}{MaxVal}, \quad (2)$$
and multiplying the feature map with this scaling factor. The dynamic range after map normalization operation is [0,255], for all feature maps.

The last thing to do is to combine all feature maps belonging to one feature into one integrated map. All maps are re-scaled into the same size, and a pixel-by-pixel maximum operation is performed among all these re-scaled maps. This resembles a winner-take-all competition among different feature maps. An additional maximum operation is performed to combine Cb and Cr feature maps into one color saliency map. Thus the computation of low-level features yields two saliency maps, i.e., the intensity saliency map and the color saliency map.

Medium-level feature (motion): In our thought, motion presents not only bottom-up but also top-down attention information. Some basic observations can be drawn:

a) Image segments with spatially consistent motion field are more likely to be parts of foreground moving objects and receive more user attentions than those in the background do.

b) The user is more aware of objects with temporally consistent motions.

c) Objects with larger motion draw more attentions than those with smaller motion do.

d) People can pay attention to a very limited number of objects in a scene. When there are many different objects (possibly with different motions), people loss the ability of attention.

In this paper, block motion vectors are used for motion analysis and user-attention modeling. Though sometimes motion vectors don't reflect the "true" motion field well, but utilizing motion vectors can greatly help for reducing more computation complexity than that of the fine-grained optical flows. An approach similar to [7] is adopted to deal with our motion saliency map.

For one macroblock *i*, we define the intensity of its motion vector (dx, dy) as
$$I(i) = \sqrt{dx^2 + dy^2}. \quad (3)$$
By computing motion vector intensities of all macroblocks in a frame, we get an intensity map I. Let $W_s(i)$ be the set of motion vectors of all macroblocks until a redefined window, W_s. The phase of one motion vector (dx, dy) is defined as
$$Phase = \arctan(\frac{dy}{dx}). \quad (4)$$
The range of a phase is $[0, 2\pi]$. We calculate an eight-bin phase histogram of $W_s(i)$ and measure the spatial-temporal consistency as
$$C(i) = 1 + \frac{\sum_{h=1}^{H_s} p_s(h)\log(p_s(h))}{\log(H_s)}, \quad (5)$$

where H_s is the number of histogram bins and $p_s(h)$ is the probability of one particular bin h. The larger $C(i)$ means the more consistent motion field. By computing the spatial-temporal consistency values of all macroblocks, we get a motion consistency map *C*.

The motion intensity map, *I*, and the motion consistency map, *C*, are combined to yield a single value of motion attention. That is,

$$M = I \cdot C. \quad (6)$$

We call *M* as the motion saliency map, and a normalization operation is performed to yield a dynamic range of [0, 255].

High-level feature (face): Dominant faces in video frames certainly attract user's attention. In fact, we can assume that people naturally locate faces in video frames with priority over other types of objects. In our implementation, two kinds of face detection schemes are investigated.

Traditional face detection is based on template matching. We investigate an object detector specifically trained for face detection. The idea has been initially proposed by Paul Viola [9] and improved by Rainer Lienhart [10]. This scheme provides comparably good results for frontal faces; however, it suffers from non-frontal faces and tilted faces.

The alternative class of face detection is based on skin-region detection. We implemented the skin color model given in [11]. This scheme can find most of face regions in video frames. While miss detection rate is very low, using skin-color detection only sometimes suffers from high false alarm rates. It is our experience that using the morphological opening operation and imposing some size and aspect ratio constraints on the detected regions can help to reduce false alarms.

Finally, after face region detection, we get a face saliency map.

2.2 The Fusion Stage and the Focus Point Detection

The fusion stage: After four saliency maps (intensity, color, motion and face maps) are available, a fusion stage is required to integrate all maps into a final saliency map. In our implementation, we use a particular fusion scheme, called priority-based competition. Feature maps in the same level are combined using the maximal operation. Then the integrated map in the lower level is scaled down by a pre-defined factor, and then competes with the integrated maps in the higher level. The block diagram of the priority-based competition is shown in Fig. 2.

Focus point detection: Once the final saliency map is generated through the fusion stage, we are now ready to detect a single focus point for each video frame. Following steps accomplish this work:

a) Thresholding and binarization. Let *MIN* and *MAX* be the minimal and the maximal values of the final saliency map. The threshold *BinThr* is determined as:

$$BinThr = MAX - (MAX - MIN) / ThrFactor, \quad (7)$$

where *ThrFactor* is an adjustable parameter.

b) Connected component analysis is then performed and the largest component is found. The center of the largest connected component is set as the candidate of focus point. If no connected component is found, or the area of the largest component is smaller than a pre-defined threshold, the candidate focus point is set to be the center of the frame.

c) To avoid possible false alarm, we restrict the distance between focus points of two neighboring frames.

d) To maintain a smooth locus of focus points, a Gaussian-like filter is applied to successive detected focuses.

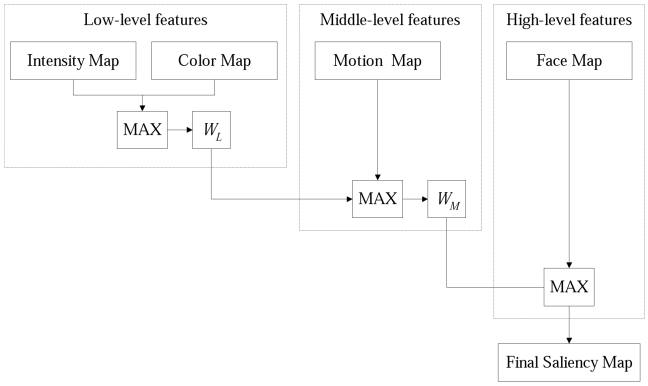

Figure 2. The priority-based fusion scheme for final saliency map calculation.

3. User-Assisted Spatial Resolution Reduction

Spatial resolution reduction is necessary for some content repurposing related applications. For example, adapting videos with higher spatial resolution to devices with smaller display. The easiest way to perform spatial resolution reduction is through direct sub-sampling, however, this may be undesirable because the interesting subject(s) may be too small to view. For better user satisfaction of spatial resolution reduction, we propose the so-called user-assisted spatial resolution reduction. First, we perform the attention focus detection based on the prescribed user-attention modeling system. Then we let the user specify the wanted spatial sub-sampling factor and the spatial cropping factor. Let these two values be respectively denoted as r_s and r_c, and the valid range of them is (0,1). Finally, we perform the required sub-sampling and cropping operations by setting the detected focus point as the center of the operational region. Then we can adjust the spatial resolution of image or video by

cropping the operational regions instead of just down sampling the entire image or video.

4. Experimental Results

In this section, we present some experimental results of the proposed user-attention based focus point detection framework, the user-assisted spatial resolution reduction, and the attentional focus-point subjective experiment.

We test the proposed focus detection framework by using some well-known sequences. Fig. 4 and Fig. 5 show some focus detection results for the sequences "foreman" and "mobile", respectively. In Fig. 4, the foreman has conspicuous motion activities and explicit faces, which are successfully detected through motion and face saliency calculations, and the detected focus point performs reasonably and satisfactorily.

(a) (b)

Figure 4. Results of saliency maps and focus point detection for the sequence "foreman": (a) the 9th frame, (b) the 80th frame.

Fig. 5 shows saliency map and detected focus points for the interested frames of the mobile sequence. Although these frames have a lot of intensity or color saliency spots (it can be seen that the background is complex), motion maps are popped out because (1) there is only one conspicuous region exists and (2) motion maps are with larger weighting in the fusion stage. The results are satisfactory.

(a) (b)

Figure 5. Results of saliency maps and focus point detection for the sequence "mobile": (a) the 15th frame, and (b) the 73rd frame.

To further validate the proposed framework of focus finding, we performed subjective experiments. Testing video clips are classified into six categories according to different characteristics. In different categories, we have different weighting factors in combining saliency maps of the focus detection framework. Then, we invited 20 observers to participate in the test. Every observer gives a score from 1 to 5 (larger value means better quality in perception) according to his or her intuition. The subjective results are shown in Table 1. In each category, the score mean is satisfactory and the score variance is small. We find from that by choosing proper parameter sets, the detected focus point in different kinds of video content is representative. In conclusion, the focus-point detection framework successfully models the observer's attention.

Category	Score mean	Score var.
Home Video	3.92	0.15
High Motion	3.81	0.09
Nature	3.43	0.01
Sports	3.94	0.21
TV & Movie	3.70	0.05
Others	3.52	0.05

Table 1. Results of the focus-point subjective experiment

The second kind of our experiments is about spatial resolution reduction. Fig. 6 shows different spatial resolution reduction results of the sequence "horn" by using different parameters but with the same final image size. The cropping regions shown in Fig. 6 are determined by the assistance of the user attention modeling system. The difference of semantic information revealed by different spatial resolution reduction parameters is obvious to see, and the user-attention modeling system do help for revealing more semantic information when the final image size is very small as compared to the original one.

(a) (b) (c)

Figure 6. Spatial resolution reduction with parameters: (a) r_s = 0.25 and r_c = 1.0, (b) r_s = 0.5 and r_c = 0.5, and (c) r_s = 1.0 and r_c = 0.25.

5. The Conclusion and Future Works

The following conclusions can be made for the proposed framework of focus detection:

a) A general user attention based focus detection framework is developed to capture users' focus points on video frames. The proposed framework considers both bottom-up and top-down attentions, and is extensible and flexible for integrating more features with a variety of fusion schemes.

b) Combining other perceptual models with our framework, the system can have many applications, such as user-attention based video encoding, and the user-assisted spatial resolution reduction, which have also been addressed in this write-up.

In the future, more tasks must be done to enlarge the capability of the proposed framework, e.g., using more complex modeling schemes to improve adopted features, integrating more robust face detection schemes and adopting more complex fusion schemes to the fusion stage.

As for applications of focus detection, one interesting application is the so-called user-attention based video encoding, which aims at reducing bitrate requirements, without sacrificing perceived quality for typical encoding schemes. The user-attention based video encoding can be done through discarding unimportant visual information as much as possible, under the guideline of the foveation model [12]. It also selectively preserves higher quality for those focused regions, in trade of worse quality for those periphery regions, to maximally match users' expectation. These scenarios can also be applied to generate the base layer bitstream of a scalable video, when the bitrate constraint is very strict. Fig 3 shows the proposed architecture of the user-attention based video encoding. Some research issues and experimental results are addressed in [13].

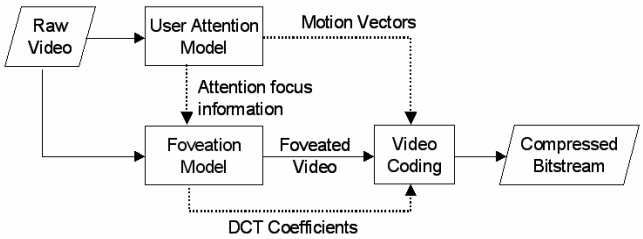

Figure 3. The proposed architecture of the user-attention based video encoding.

References

[1] A. M. Treisman and G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, Vol. 12, No. 1, pp. 97-136, 1980.

[2] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, Vol. 4, pp. 219-227, 1985.

[3] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Reviews Neuroscience*, Vol. 2, No. 3, pp. 194-203, Mar. 2001.

[4] K. Schill, E. Umkehrer, S. Beinlich, G. Krieger, and C. Zetzsche, "Scene analysis with saccadic eye movements: top-down and bottom-up modeling," *Journal of Electronic Imaging, Special Issue on Human Vision and Electronic Imaging*, Vol. 10, No. 1, pp. 152-160, 2001.

[5] I. A. Rybak, V. I. Gusakova, A. Golovan, L. N. Podladchikova, and N. A. Shevtsova, "A model of attention-guided visual perception and recognition," *Vision Resolution*, Vol. 38, pp. 2387-2400, 1998.

[6] G. Deco and J. Zihl, "A neurodynamical model of visual attention: Feedback enhancement of spatial resolution in a hierarchical system," *Journal of Computational Neuroscience*, Vol. 10, pp. 231-253, 2001.

[7] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. ACM Multimedia* (ACMMM'02), pp. 533-542, Dec. 2002.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 20, No. 11, pp. 1254-1259, Nov. 1998.

[9] P. Viola and M. Jones, "Robust real-time object detection," in Second Intl, Workshop on Statistical and Computational Theories of Vision: Modeling, Learning, Computing and Sampling, July 2001.

[10] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in Proc. IEEE Intl. Conf. Image Processing (ICIP'02), pp. 900 – 903, Sept. 2002.

[11] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," IEEE Trans. Multimedia, Vol. 1, No. 3, pp. 264 – 277, Sept. 1999.

[12] C.-C. Ho and J.-L. Wu, "A foveation-based rate shaping mechanism for MPEG videos," in *Proc. 3th IEEE Pacific-Rim Conference on Multimedia* (PCM'02), Springer-Verlag (LNCS 2532), pp. 485-492, Hsinchu, Taiwan, Dec. 2002.

[13] Ho, C.-C. "A Study of Effective Techniques for User-Centric Video Streaming", Ph.D. dissertation, National Taiwan University, Taipei, Taiwan, June, 2003.