A New Hybrid Model for Video Shot Saliency Extraction

Tao Fang
College of Fundamental Education
Sichuan Normal University
ChengDu 610066, China
f_fangtao@163.com

***Abstrac*--Video screen shots of images are special, and they are motive compare to other images. To extract saliency maps from video images quickly and accurately is a hard task. This paper is inspired by some excellent works, employing the combination of several ideas to propose a hybrid model for extracting saliency. The hybrid model to extract salient region is based on visual attention theory, dynamic feature saliency maps extraction technique, and motion-prior idea integrate spatial with time. We perform experiments to make extraction with itti model, and results prove the fact that the proposed model can get an ideal effect when extract salient region dynamically considering the combination of speed and direction is closing to biological visual mechanisms.**

Keyword:--visual attention; saliency extraction; video image.

I. INTRODUCTION

One of the key technologies required for efficient recognition of video data is saliency extraction. A concise and informative video saliency extraction can be used widely spread to various areas, such as robotic systems, and traffic monitoring.

The nature of video image is unsuitable for traditional forms of data process, retrieval, and saliency detection. All the differences make the video shot data employ a targeted Saliency detection method, by which computational mount, computational speed, and effect can be take in to consideration. Saliency detection or extraction methods are broadly used in many fields, which can broadly be classified as biologically based, purely computational, or a combination. In general, all the methods employ a low-level approach by determining a contrast of image regions relative to their surroundings, using a signal or combined features, such as intensity, color, and orientation.

A classic biologically method is the visual attention, which is proposed by shausen et al. [1] and worked by Itti et al[2],[3]. Itti and his partners determine center-surround contrast using a Difference of Gussians (DoG) approach, which is used widely. A method inspired by Itti was proposed in [4], they compute center-surround difference with square filters and use integral images to speed up calculations. However, most existing visual attention approaches are based on the bottom-up computational framework [5],[6],which are not fit for the video shot data perfectly.

Some methods represented in [7],[8],[9],[10] are purely computational. Zhang, Ma [7] and Achanta et al.[8] estimate saliency using center-surround feature distances. Hu et al. [9] estimate saliency by a frequency-tuned method. A remarkable computational method using motion features was given by Chen[10] , by which motion features can be extracted from some dynamic images.

The third category methods combined bio-models and computational ones. These methods mostly targeted a fixed image type, though which can get an ideal result fit for a specific situation. However, many methods combined the visual attention model with other ideas [11]. Researchers propose a novel saliency detection algorithm based on the idea of maximum symmetric surround, trying to exploits features of color and luminance is simple to implement and is computationally efficient [19]. Radhakrishna[20] proposes a method that outputs full resolution saliency maps with well-defined boundaries of salient object. The boundaries are preserved through retaining more substantially frequent content from the original image.

We aims to find a hybrid method can fit for the video shot, at the same time, computational mount, applicability, speed, notable effect must be taken into consideration.

This paper organized as follows. Section 2 summarizes some facts of image saliency like visual attention mechanism, dynamic saliency extraction based on motion features. Section 3 presents our saliency extraction method which combined the visual attention model with dynamic saliency extraction for video. Some experiment results and the final conclusion are given in Section 4 and 5.

II. PRELIMINARY

This section summarizes itti method, the saliency map extraction based on the visual attentions mechanism, and reviews the saliency map extraction method based on motion features.

A. Visual Attention model

Visual Attention model is drawn from neurobiological conception. This model allows us to break down the problem of understanding a live video sequence into a series of computationally less demanding and localized visual, audio, and linguistic analytical problems. As some research, visual objects with special attributes attracting observers' attention then we can say the saliency is produced.

Objects selection of human eyes enables us to pin out the areas attracting us in the dynamic video situations. Visual Attention is a mechanism for handling information, which can be used in saliency extraction.

[1] proposed the model for the first time. This paper presented a biological plausible model of an attentional mechanism for forming position and scale-invariant representations of objects in the visual world. This model is widely accept by[2]. Authors utilize the model to get the most stimulating parts, and use grayscale image to represents degree of saliency.

B. Itti Model

Itti is classical model based on visual attention model. In visual attention models, the attention point is the Maxima of the saliency maps, attention regions are the round areas centered the attention points with fixed radius. That means human can find meaningful information within enormous data with their eyes. Visual attention tells us that saliency can give rise to attention. Visual attention model describes the attention with grayscale images. Generally, the bigger salient value we get and the more attention achieve.

Most visual approaches can be divided into three steps [12]: The first step is feature extraction in which multiple low-level visual features such as intensity, color, orientation, texture, motion, are extracted from the image at multiple scales. The second step is saliency computation which is based on the center-surround operation, self-information or graph-based random walk using multiple features. A master map or a saliency map is computed after a normalization and linear/nonlinear combination. The last step is to identify a few key locations on saliency map with winner-take-all, or inhibition-of-return, or other nonlinear.

As given information, itti follows the three steps, which is a representative method based on visual attention mechanism. Framework map of itti[12] is shown in fig.1, which illustrate the flow when compute the saliency.

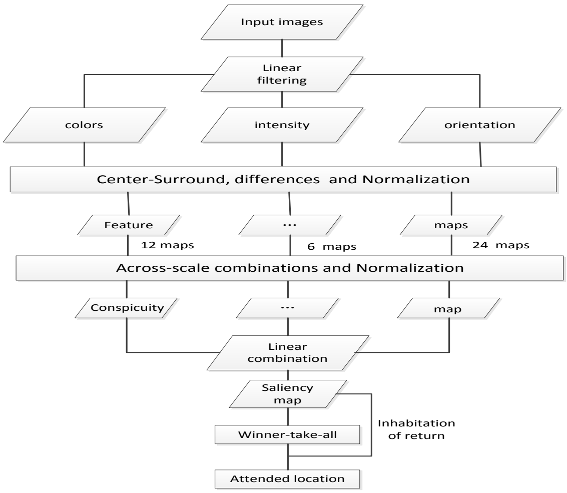

Figure 1. Itti framework map

As shown in Figure 1, at the first move, features such as colors, orientations, intensity are extracted from initial input map though Gussian linear filter. Then, 12 color maps, 6 intensity maps, 24 orientation maps are formed though the center-surround and normalization computation. These maps must be processed though further standard operations, mathematical function N( ) for short to take the computational task. The next thing is to get the conspicuity maps though normalization and feature maps combination. Conspicuity maps can be tuned into salient maps though linear combination.

Observer will get the salient regions though two-level Neural Network winner-take-all handling the salient maps. With the inhabitation of return, current salient region is inhibited to search the next salient region. Follow the Figure 1, researchers will finally get the salient regions of specified images.

However, video image is changing rapidly which means we have to change a static feature correspondingly. Considering the situation where the video shots complex cases. Then the static extraction is ineffective. How about we use a feature set though which to extract saliency introduced by [14]. The paper proposes a set of novel features, including multiscale contrast, center surround histogram to describe a salient object. The method is feasible but they can't work in the video saliency extraction because the motion features in this context.

C. Hybrid Model

As a primary property of video shot, both dynamic and static content reflects Saliency dependent on different situations. The hybrid model is an adaptive motion-time dynamic hybrid model inspired by the motion-prior idea given by [13]. This model is based on the fact of the visual character, to combine the dynamic saliency maps and static saliency maps. the model can be illustrated by the following formula.

$$A = \sum_{i,j \in F} (W_s \times I_{i,j}^{s} + W_t \times I_{i,j}^{t}) \tag{3}$$

W_s, W_t represent the Weights of spatial and time respectively. $I_{i,j}^{s}$ is the spatial saliency map and $I_{i,j}^{t}$ is the time saliency map. The situation of the video shot is unpredictable, we can't tell what will happen next frame but we can tell which is the salient region, neither do the computer in the dynamic case of video shots. If we use time saliency map only, we may lost some motion features. The same thing will occur when we use spatial saliency map only.

It is hard to suit to complex applications if the Weights are fixed. Hence, we employ an adaptive method that the Weights can be revised automatically; the method is motion-prior hybrid model combining the dynamic salient degree with static salient degree. When motive contrast varies, the model can adjust the two Weights' proportion. Fig.2 show the Weights' proportion's changing discipline.

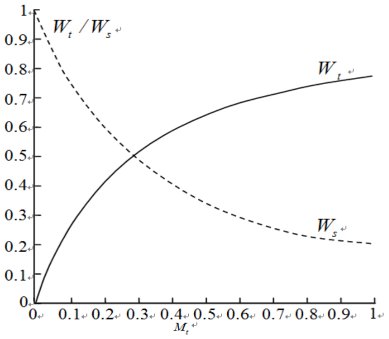

Figure 2. Changing discipline of Weights' proportion

Compare to previous methods, this model incarnates the proportion between dynamic areas and static areas. When it works, it resist to low texture and noise. Even in terrible video resolution, we can also to detect saliency areas.

III. Hybrid Model with motion direction

Our work is to propose a hybrid model inspired by hybrid model. In this model, we bring motion direction as a feature combing it into our model. As introduction in [11], speed and movement directions will attract great attention, one reason is that human is sensitive to contrast thing, especially human's visual system. The greater visual contrast, the more visual saliency will get. That means the motion in the video can also generate saliency. Utilizing the visual contrast to compute saliency maps is a reasonable method for the video shot images. According to sum of absolute difference rule, we can get the contrast degree and the formula as following.

$$SAD = \sum_{x=1}^{M} \sum_{Y=1}^{N} | F_K(x,y) - F_{(K-1)}(x+u, y+u) |$$

To compute motion vector, which reflect change rate of video image, we can get the motion feature maps to describe the speed and direction of the motive object. To detect diversification of video shot on different direction is an effective way to get feature maps, motion orientation conspicuity maps (MOCM). Motion saliency detection is widely used, such as [15]. In this paper, author proposed a method to get saliency detection on video slices though separate foreground motion objects from backgrounds. Motion is one salient feature in our model.

The extraction method based on motion feature also combined with some classic ideas, such as center-surround [16], difference computation etc. However, during the combination of motion features, we try a wavelet mergence to get saliency maps. Then, the whole framework is completed and Figure2 shows the motion feature extraction flow.

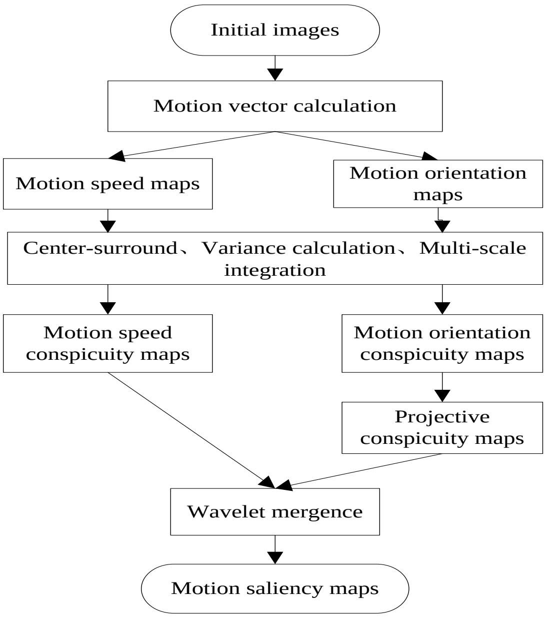
Initial images
Motion vector calculation
Motion speed maps
Motion orientation maps
Center-surround、 Variance calculation、 Multi-scale integration
Motion speed conspicuity maps
Motion orientation conspicuity maps
Projective conspicuity maps
Wavelet mergence
Motion saliency maps

Figure 3. Motion saliency extraction

Motion features can be projected into 4 directions, 0°,45°,90°and 135°to form 4 motive direction maps. Each of the maps reflects the displacement between different adjacent frames. When we get various displacements which means frames are motive in different directions. Observers from speed maps though displacements, which demonstrate the speed rate changing of pixels. We construct Gaussian pyramid for both maps to extract motion saliency maps for different resolution. Then we can obtain motion feature of every pixel in new map though specific algorithm. Using the following formula, observer computes the motive distance of every pixel between two adjacent frames.

$$M_s = \sqrt{u'(i,j)^2 + v'(i,j)^2} \tag{1}$$

M_s is pixel movement distance between two frames. $u^{'}(i,j)$ represents displacement on the 0 ° directions and $v^{'}(i,j)$ represents displacement project to 0 ° direction from 90°direction. Then we can get four projections on various directions. $M_{0\_0}(i,j)$, $M_{0\_45}(i,j)$, $M_{0\_90}(i,j)$ and $M_{0\_135}(i,j)$ represent these projections. Put them into Gaussian pyramid low pass filter and center-rounded, the formulas shows as following

$$M_s(c,s) = \left|M_s(c)\Theta M_s(s)\right|$$

$$M_{0\_0}(c,s) = \left|M_{0\_0}(c)\Theta M_{0\_0}(s)\right|$$

$$M_{0\_45}(c,s) = \left|M_{0\_45}(c)\Theta M_{0\_45}(s)\right|$$

(2)

$$M_{0\_90}(c,s) = \left|M_{0\_90}(c)\Theta M_{0\_90}(s)\right|$$

$$M_{0\_135}(c,s) = \left|M_{0\_135}(c)\Theta M_{0\_135}(s)\right|$$

We compute these formulas at projective conspicuity maps and c is the central scale, s is the marginal scale and $c \in \{2,3,4\}$. Θ is difference operation between images with different scales. M_i denotes five feature maps. We standardize these maps and merge them to obtain speed saliency map and motion direction map. Extract saliency map with wavelet technology subsequently. Follow the Fig.2, we can get an ideal saliency maps according to what we see from the video shot.

A remarkable example [11] is given in Fig.3, from which we can see three boats traveling on the sea. We can pick up some key frames which contains the most useful information. These frames can be processed into a feature map on projective coordinate with center-surround algorithm. The truth is found that the most distinct motive contrast is at the 0°with respect to the motive direction is horizontal, and nothing happens at the 90°,so we get a map is empty except the black background.

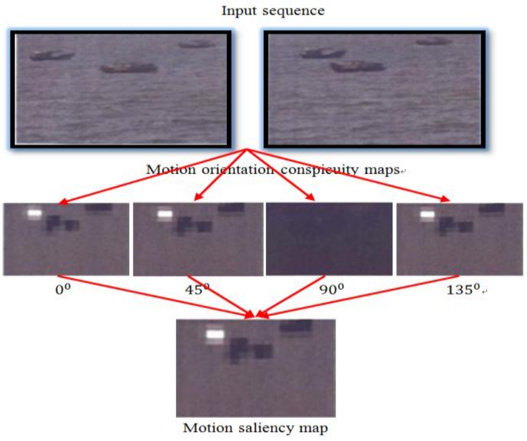

Figure 4. Motion feature saliency map

Fig 4 shows the result of saliency map with motion direction. As we can see, this is a feasible method to detect saliency map in different views and the effect is significant. One can't get a valid map without orientation detection in shown situation.

IV. EXPERIMENT RESULT

To illustrate the effectiveness of our method, we test the proposed method on typical video selected from internet randomly. This video shows the three people pass by with high speed on skating. Experiment is shown in Fig.5. In this picture, three men are skating on the road and there is a striking sign with speed limited 40. Two adjacent images are quite different because three men skate away, but the sign still be there.

If we use motion feature only, then we may lost the sign in salient regions. Correspondingly, if we employ itti model only, we may lose the salient regions of three men. We use our model to handle input video shots, as we can see in Fig5 a better effect made, and more close to the bio-visual mechanism, which can capture the saliency region accurately.

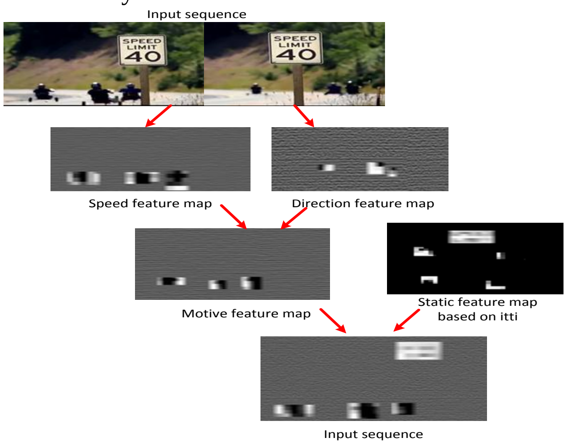

Figure 5. Experiment result

This method works well for dynamic video shots. For those videos with fast motions, e.g. sports videos s, the proposed method cannot handle them directly. For these videos, we need to do shot detection first, and then perform the proposed method on them except close-up shots.

Our method quite efficiency and efficacy, we record the execution time during the experiments. Table 1 shows the execution time of key parts of the methods.

TABLE 1 EXECUTION TIME OF KEY PROCESS

Process	Execution time
Speed projection	0.13secs
Direction projection	0.18secs

Execution time is table 1 is the mean time of hounds of experiments. We can promote the efficiency in future work.

V. COMPARISON

Itti model is the first one to merge visual feature into saliency map. Then the model employs a dynamic natural network to extract motion regions, which is suitable for static picture. Drawback of the model is that the saliency regions mismatch with human visual mechanism.

[17] proposes a method is based on detecting salient object, but it is too simple to satisfy practical operation and it get the same result with the itti model. [18] proposes an Gaussian Mixture Model, which is widely used in stable situation. This model uses several Gaussian functions to fit gray distribution. Drawback of this model is fragile to noise, when there is more noise in video and result would be awful. Our hybrid model absorb in motion directions to detect salient regions. From experiment, we can see that the proposed mode in the paper works well in video shot saliency extraction.

Our model based on hybrid model, but the key processes are efficient. Experiment results show the feasibility of the proposed model.

VI. CONCLUSION

This model we present here shows that combination of several methods can get an ideal result at specific situation, especially fit for the video shot image data. Our model is based on some excellent works. We utilize the motion direction as a feature. Combine hybrid model which can adjust the important parameters according to the motive features. That the low computational mount and not to be affected by the low level veins are the advantages. However, a general model fit for various circumstances have still to be researched.

[1] B. Olshausen, C. Anderson, and D. Van Essen.A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. Journal ofNeuroscience, 13:4700–4719, 1993

[2] L. Itti, C. Koch, and E. Niebur.A model of saliency-basedvisual attention for rapid scene analysis. IEEE Transactionson Pattern Analysis and Machine Intelligence, 20(11):1254–1259, 1998.

[3] Itti L., Koch C. Computational Modeling of Visual Attention[J]. Nature ReviewsNeuroscience,2001,2(3):194–203.

[4] S. Frintrop, M. Klodt, and E. Rome.A real-time visual attentionsystem using integral images.International Conferenceon Computer Vision Systems, 2007.

[5] N.Bruce and J. Tsotsos, Saliency Based on information Maxmization. Advances in Neural Information Processing System, pp.155-162, MIT Press, 2005.

[6] J.Harel, C.Koch, and P. Perona , Graph-Based visual Saliency Advances in Neural Information Processing System, pp.545-552, MIT Press, 2006.

[7] Y.-F. Ma and H.-J.Zhang.Contrast-based image attentionanalysis by using fuzzy growing.In ACM International Conferenceon Multimedia, 2003.

[8] R. Achanta, F. Estrada, P. Wils, and S. S¨usstrunk.Salientregion detection and segmentation.International Conferenceon Computer Vision Systems, 2008.

[9] Y. Hu, X. Xie, W.-Y. Ma, L.-T.Chia, and D. Rajan. Salientregion detection using weighted feature maps based on thehuman visual attention model.Pacific Rim Conference onMultimedia, 2004.

[10] Jiawei Chen. Research and application for visual attention computational model [D].Xia men university,2009.

[11] J. Harel, C. Koch, and P. Perona.Graph-based visual saliency. Advances in Neural Information Processing Systems,19:545–552, 2007.

[12] Peng Jiang, Xiao lin Qin. Efficient visual attention region detection in Dynamic Scene. Journal of Chinese Computer System.4(21),pp:163-167.2010.

[13] MilanceseR,GilS,et al. Attentive mechanisms for dynamic and static scene analysis[J].Opt Eng, 1995, 34(8):2428-2434

[14] Liu T, Yuan Z, Sun J, et al. Learning to detect a salient object[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2011, 33(2): 353-367.

[15] Cui X, Liu Q, Metaxas D. Temporal spectral residual: fast motion saliency detection[C]//Proceedings of the 17th ACM international conference on Multimedia. ACM, 2009: 617-620.

[16] Klein D A, Frintrop S. Center-surround divergence of feature statistics for salient object detection[C]//Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011: 2214-2219.

[17]Liu F, Gleicher M. Region enhanced scale-invariant saliency detection[C]//Multimedia and Expo, 2006 IEEE International Conference on. IEEE, 2006: 1477-1480.

[18] M. Haque, M. Murshed and M. Paul, A hybrid object detection technique from dynamic background using Gaussian mixture models, IEEE 10th Workshop on Multimedia Signal Processing, 2008.

[19]Achanta R, Susstrunk S. Saliency detection using maximum symmetric surround[C]//Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE, 2010: 2653-2656.

[20]Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 1597-1604.