Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling

Sheng-hua Zhong[1], Yan Liu[1], Feifei Ren[1,2], Jinghuan Zhang[2], Tongwei Ren[3]
[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, P.R. China
[2]School of Psychology, Shandong Normal University, Jinan, P.R. China
[3]Software Institute, Nanjing University, Nanjing, P.R. China
{csshzhong,csyliu,csffren}@comp.polyu.edu.hk, xlxy@sdnu.edu.cn, rentw@software.nju.edu.cn

Abstract

Human vision system actively seeks salient regions and movements in video sequences to reduce the search effort. Modeling computational visual saliency map provides important information for semantic understanding in many real world applications. In this paper, we propose a novel video saliency detection model for detecting the attended regions that correspond to both interesting objects and dominant motions in video sequences. In spatial saliency map, we inherit the classical bottom-up spatial saliency map. In temporal saliency map, a novel optical flow model is proposed based on the dynamic consistency of motion. The spatial and the temporal saliency maps are constructed and further fused together to create a novel attention model. The proposed attention model is evaluated on three video datasets. Empirical validations demonstrate the salient regions detected by our dynamic consistent saliency map highlight the interesting objects effectively and efficiency. More importantly, the automatically video attended regions detected by proposed attention model are consistent with the ground truth saliency maps of eye movement data.

Introduction

As a cognitive process of selectively concentrating on one aspect of the environment while ignoring other things, attention has been referred to as the allocation of processing resources (Anderson 2004). The strongest attractors of attention are stimuli that pop-out from their neighbors in space or time usually referred to as "saliency" or "novelty" (Itti & Baldi 2009). Visual attention analysis simulates the human vision system behavior by automatically producing saliency maps of the target image or video sequence (Zhai & Shah 2006). This saliency map is proposed to measure of conspicuity and calculate the likelihood of a location in visual data to attract attention (Koch & Ullman 1985).

Owing to the visual saliency detection provides predictions about which regions are likely to attract observers' attention (Parkhurst et al. 2002), it has a wide range of applications. These applications include image/video representation (Wang et al. 2007), object detection and recognition (Yu et al. 2010), activity analysis (Oikonomopoulos et al. 2006), content aware resizing (Avidan & Shamir 2007), object tracking (Yilmaz et al. 2006), and robotics controls (Jiang & Crookes 2012).

The saliency of a spatial location depends mainly on two factors: one is task independent and the other is task dependent (Marat et al. 2009). The first one is called bottom-up which is mainly depending on the intrinsic features of the visual stimuli. The latter refers to top-down process which integrates high-level information and cognitive understanding. Most existing computational visual saliency models follow a bottom-up framework that generates independent saliency map in each selected visual feature space (Li et al. 2012). After feature maps are computed, they are normalized and combined into a master saliency map that represents the saliency of each pixel (Koch & Ullman 1985; Itti et al. 1998).

Video saliency detection is to calculate the salient degree of each location by comparing with its neighbors both in spatial and in temporal areas (Li et al. 2012). In these years, image attention detection has been long studied, while not much work has been extended to video sequences where motion plays an important role. Neurophysiological experiments have proved that neurons in the middle temporal visual area (MT) compute local motion contrast. And such neurons underlie the perception of motion pop-out and figure-ground segmentation which influences the attention allocation (Born et al. 2000). After realizing the importance of motion information in video attention, the motion feature has been added into the saliency models (Cheng et al. 2005; Peters & Itti 2008). Recently, to simulate two pathways (magnocellular and parvocellular) of the human visual system, the video saliency detection procedure are divided into spatial and temporal pathways (Marat et al. 2009). These two pathways correspond to the static and dynamic information of video.

In existing video saliency detection models, optical flow is the most widely used motion detection approaches. These models rely on the classical optical flow method to extract the motion vector between each frame pair independently as the temporal saliency map (Xu et al. 2010; Mathe & Sminchisescu 2012; Loy et al. 2012). They integrate the temporal saliency map with different spatial saliency map to construct entire attention model. Unfortunately, although the optical flow technique can accurately detect motion in the direction of intensity gradient, the temporal saliency is not perfectly equal to the amplitude of all the motion between each adjacent frame pair. Indeed, only

the continuous motion of the prominent object with enough amplitude can be popped out as the indicator of temporal salient region. In addition, the independent calculation of each frame pair leads to a high computational complexity.

To address the problem in existing saliency detection methods, this paper proposes a novel spatio-temporal attention model (STA) by referencing the characters of the human vision system. In spatial saliency map, we follow the procedure of the classical bottom-up spatial saliency map based on low-level features. In temporal saliency map, a novel dynamic consistent optical flow model (DCOF) is proposed based on the human visual dynamic continuity. Different from the classical optical flow model estimates motion between each adjacent frame pair independently, the proposed DCOF takes account of the motion consistency in the current frame and between consecutive frames.

In the following parts of this paper, we first discuss the related work of optical flow methods. Then, a novel spatio-temporal video saliency detection technique is introduced. In the experiment part, we demonstrate the performance of the proposed attention model on three video sequence datasets. The paper is closed with conclusion.

Related Work on Optical Flow

The concept of optical flow was first studied in the 1940s and ultimately published by psychologist James J. Gibson (1950). It is defined as the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and the scene (Warren & Strelow 1985). Optical flow approach approximates the object motion by estimating vectors originating or terminating at pixels in image sequences, so it represents the velocity field which warps one image into another feature space (Liu et al. 2009). The motion detection methods based on optical flow technique can accurately detect motion in the direction of intensity gradient.

The classical formulation of optical flow is introduced by Horn and Schunck (1981). They optimize a functional based on residuals from the brightness constancy constraint, and a regularization term expressing the smoothness assumption of the flow field. Black and Anandan addressed the outlier sensitivity problem of HS model by replacing the quadratic error function with a robust formulation (Black & Anandan 1996). Here, we refer to all these formulations which are directly derived from HS as the "classical optical flow model." The objective function of the classical optical flow is defined as:

$$\begin{aligned} E(\mathbf{u},\mathbf{v}) = \sum_{i,j} \{ & f_D(I_1(i,j) - I_2(i+u_{i,j}, j+v_{i,j})) \\ & + \lambda[f_S(u_{i,j} - u_{i+1,j}) + f_S(u_{i,j} - u_{i,j+1}) \\ & + f_S(v_{i,j} - v_{i+1,j}) + f_S(v_{i,j} - v_{i,j+1})]\} \end{aligned} \quad (1)$$

where $\mathbf{u}$ and $\mathbf{v}$ are the horizontal and vertical components of the optical flow field to be estimated from image I_1 and I_2. λ is a regularization parameter. f_D is the brightness constancy constraint function, and f_S is the smooth penalty function.

Although different efforts have been put into improving the optical flow, the median filtering the intermediate flow results after each warping iteration (Wedel et al. 2008) is the most important source to improve the performance of classical model. According to the extensive test by (Sun et al. 2010), the median filtering makes non-robust methods more robust and improves the accuracy of all optical flow models. The optimization of Eq. (1), with interleaved median filtering, can be approximately minimized as Eq. (2) (Li & Osher 2009; Sun et al. 2010):

$$\begin{aligned} E(\mathbf{u},\mathbf{v},\hat{\mathbf{u}},\hat{\mathbf{v}}) = \sum_{i,j} \{ & f_D(I_1(i,j) - I_2(i+u_{i,j}, j+v_{i,j})) \\ & + \lambda[f_S(u_{i,j} - u_{i+1,j}) + f_S(u_{i,j} - u_{i,j+1}) \\ & + f_S(v_{i,j} - v_{i+1,j}) + f_S(v_{i,j} - v_{i,j+1})]\} \\ & + \lambda_2(\|\mathbf{u} - \hat{\mathbf{u}}\| + \|\mathbf{v} - \hat{\mathbf{v}}\|) \\ & + \sum_{i,j} \sum_{(i^\dagger, j^\dagger) \in N_{i,j}} \lambda_3(|\hat{u}_{i,j} - \hat{u}_{i^\dagger, j^\dagger}| + |\hat{u}_{i,j} - \hat{u}_{i^\dagger, j^\dagger}|) \end{aligned} \quad (2)$$

Here, $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ denote an auxiliary flow field. $N_{i,j}$ is the set of neighbors of pixel (i,j). λ_2 and λ_3 are scalar weights. The last term in this equation imposes a smoothness assumption within a region corresponding to the auxiliary flow field.

Taking the state similarity into consideration, the last term of the "improved" optimal function in Eq. (2) can be updated to be Eq. (3), where the weight $\omega_{i,j,i^\dagger,j^\dagger}$ is often defined based on the spatial distance, intensity distance, and the occlusion state.

$$\sum_{i,j} \sum_{(i^\dagger, j^\dagger) \in N_{i,j}} \omega_{i,j,i^\dagger,j^\dagger} (|\hat{u}_{i,j} - \hat{u}_{i^\dagger, j^\dagger}| + |\hat{u}_{i,j} - \hat{u}_{i^\dagger, j^\dagger}|) \quad (3)$$

Spatio-Temporal Attention Model

In this section, we propose a novel spatio-temporal attention technique (STA). The schematic illustration of the proposed technique STA is described in Figure 2.

The whole spatio-temporal attention model can be partitioned into two pathways. In spatial saliency map construction, we follow the procedure of the classical bottom-up spatial saliency map. In temporal saliency map, a novel dynamic consistent optical flow model is proposed based on the human visual dynamic continuity. Different from the classical optical flow model estimates motion between each adjacent frame pair independently, the proposed DCOF both underlines the consistency of motion saliency in the current frame and between the consecutive frames.

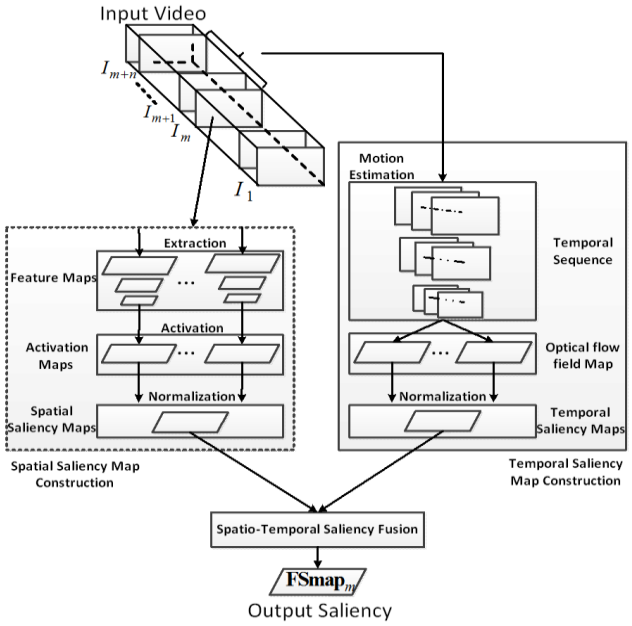


Figure 2: Schematic illustration of STA technique

Spatial Saliency Map Construction

In spatial saliency map construction, we inherit the classical bottom-up spatial saliency map based on intensity, color, contrast, and orientation features in pixel-level. The leading models of spatial saliency map construction can be organized into three stages:

1) Extraction: multiple low-level visual features such as intensity, color, orientation, texture are extracted at multiple scales;
2) Activation: The activation maps are built based on multiple low-level feature maps;
3) Normalization: The saliency map is constructed by a normalized combination of the activation map.

In our model, the graph-based saliency (GBVS) method is utilized to construct the spatial saliency map (Harel et al. 2007). In GBVS, the fully-connected directed graph G_A and G_N are constructed in activation and normalization stages separately. The weight of the directed edge in G_A from node (i, j) to node (p,q) is assigned as Eq. (4),

$$w_A((i,j),(p,q)) \triangleq d((i,j)\,\|\,(p,q)) \cdot G(i-p, j-q) \quad (4)$$

$$d((i,j)\,\|\,(p,q)) \triangleq \left|\log \frac{F(i,j)}{F(p,q)}\right|, G(a,b) \triangleq \exp\left(-\frac{a^2+b^2}{2{\sigma_G}^2}\right) \quad (5)$$

where $d((i,j)\,\|\,(p,q))$ is utilized to measure the dissimilarity between some region around (i, j) and (p,q) in the specified feature map F. σ_G is a free parameter of the algorithm.

The weight of the directed edge in G_N from each node (i, j) to node (p,q) is assigned as Eq. (6), where A is the activation map.

$$w_N((i,j),(p,q)) \triangleq A(p,q) \cdot G(i-p, j-q) \quad (6)$$

The spatial saliency map **SSMap** is formed by utilized the fully-connected directed graph G_A and G_N.

Temporal Saliency Map Construction

In related work of optical flow, we have introduced the "classical" and "improved" optical flow objective function. In video saliency detection, most of algorithms built the temporal saliency map based on the classical objective function in Eq. (1). Although the optical flow technique can detect motion accurately, the temporal saliency is not perfectly equal to the amplitude of all the motion between each adjacent frame pair. In fact, some subtle motions between frames are often resulted from the illumination change or unsteady small-disturbance in environment. Therefore, only the consistent motion of the prominent object with enough amplitude can be popped out as the indicator of salient region. In addition, the independent calculation in each frame pair has a high computational cost.

To address the problem due to the direct use of the classical optical flow in temporal saliency detection, we propose a novel optimal function. Based on the dynamic continuity of neighbor locations in the same frame and same locations between the neighbor frames, the objective function can be represented as Eq. (7):

$$\begin{aligned}
&\arg\min_{\mathbf{u},\mathbf{v},n} E(\mathbf{u},\mathbf{v},\hat{\mathbf{u}},\hat{\mathbf{v}}) = \sum_{i,j}\{f_D[\sum_{k\le n}(I_m(i,j) - I_{m+k}(i+ku_{i,j}, j+kv_{i,j}))] \\
&+\lambda_1[f_S(u_{i,j}-u_{i+1,j}) + f_S(u_{i,j}-u_{i,j+1}) \\
&+ f_S(v_{i,j}-v_{i+1,j}) + f_S(v_{i,j}-v_{i,j+1})]\} \\
&+\lambda_2(\|\mathbf{u}-\hat{\mathbf{u}}\| + \|\mathbf{v}-\hat{\mathbf{v}}\|) + \sum_{i,j}\sum_{(i^\dagger,j^\dagger)\in N_{i,j}} \lambda_3\left(\left|\hat{u}_{i,j}-\hat{u}_{i^\dagger,j^\dagger}\right| + \left|\hat{u}_{i,j}-\hat{u}_{i^\dagger,j^\dagger}\right|\right) \\
&s.t. \sqrt{{u_{i^*,j^*}}^2 + {v_{i^*,j^*}}^2} \le \sigma_o(i^*,j^*), (i^*,j^*) = \arg\max_{i,j}\left(\sqrt{{u_{i,j}}^2 + {v_{i,j}}^2}\right)
\end{aligned} \quad (7)$$

where I_m is the m^{th} frame in video sequence $\mathbf{X}$, n is the number of neighbor frames with consistent motion. It is constrained in a reasonable range by imposing the constraint based on the observation standard deviation of human visual system $\sigma_o(i,j)$. Given by the eccentricity scaling parameter c, $\sigma_o(i,j)$ is calculated by Eq. (8) (Carrasco & Frieder 1997; Vul et al. 2010),

$$\sigma_o(i,j) = cr + 0.42c\sqrt{i^2+j^2} \quad (8)$$

where r is denoted as the number of pixels per degree of visual angle. Finally, the temporal saliency map **TSMap** is formed based on optical flow filed map (**u**,**v**). The detailed procedure of the dynamic consistent saliency detection model DCOF is described in Algorithm 1.

Spatio-Temporal Saliency Fusion

When the static and temporal saliency maps are constructed, we fuse them to get the final video saliency map. Dif-

ferent fusions methods can be utilized, such as "mean" fusion, "max" fusion, and "multiplicative" fusion. Because "max" integration method has been indicated as best performance in the spatial-temporal integration function (Marat et al., IJCV, 2009). In this paper, we simply adopt the "max" fusion to take for the maximum of two saliency maps as following Eq. (9):

$$\mathbf{FSmap} = \max(\mathbf{SSmap}, \mathbf{TSmap}) \quad (9)$$

Algorithm 1: Dynamic Consistent Saliency Detection

Input: Video sequence data $\mathbf{X}$; Number of frames N_f; Pyramid level N_p.

Output: Temporal saliency map **TSmap**.

while $m < N_f - n$
  p=1;
  while $p < N_p + 1$ **do**
    $(\mathbf{u}, \mathbf{v}) = \arg\min_{\mathbf{u},\mathbf{v}} E(\mathbf{u}, \mathbf{v}, \hat{\mathbf{u}}, \hat{\mathbf{v}})$;
    $(i^*, j^*) = \arg\max_{i,j}(\sqrt{u_{i,j}{}^2 + v_{i,j}{}^2})$;
    if $\sqrt{u_{i^*,j^*}{}^2 + v_{i^*,j^*}{}^2} > \sigma_o(i^*, j^*) / 2^{p-1}$ **and** $n > 1$
      $n = \max(n - 1,1)$; $p = 1$;
    else
      $p = p + 1$;
    end if
  end while
  $\mathbf{TSmap}_m(i, j) = \text{normalize}(\sqrt{u_{i,j}{}^2 + v_{i,j}{}^2})$;
  $m = m + n$;
end while

Empirical Validation

Evaluation Setup

To illustrate the effectiveness of our model, in this section, we conduct three experiments for video saliency detection task. In the first experiment, the proposed dynamic saliency detection model is tested on the Hollywood2 natural dynamic human scene videos dataset (Marszałłek et al. 2010). The second experiment includes three typical News videos collected from YouTube. The third experiment is evaluated on the largest real world actions video dataset with human fixations (Mathe & Sminchisescu 2012).

We compare the performance of the proposed model with classical motion detection models and various saliency detection models, including: representative optical flow models (Horn & Schunck 1981; Black & Anandan 1996; Sun et al. 2010); existing spatial saliency map models (Itti et al. 1998; Harel et al. 2007) combined with different dynamic saliency detection models; and video saliency predictive model (Li et al. 2012).

For parameters in dynamic consistent saliency model DCOF, we follow the general setting of optical flow model. Convex Charbonnier penalty function is implemented as the penalty function. The number of warping steps per pyramid level is set as 3. The regularization parameter λ is selected as 5. Ten steps of alternating optimization as every pyramid level and change λ_2 logarithmically from 10^{-4} to 10^2. λ_3 is set as 1. And a 5×5 size rectangular window is used to match the size of the median filter. Follow the general setting in GBVS, the parameter σ_G is set as 5. Follow the general setting in (Vul et al. 2010), the eccentricity scaling parameter c is set as 0.08.

Experiments on Natural Dynamic Scene Videos

In the first experiment, we evaluate the proposed dynamic consistent saliency detection model DCOF on the Hollywood2 natural dynamic human scene videos dataset (Marszałłek et al. 2010). This dataset contains the natural dynamic samples of human in ten different natural environments, including: house, road, bedroom, car, hotel, kitchen, living room, office, restaurant, and shop. These dynamic scene video samples are acquired from Hollywood movies. In this experiment, we want to demonstrate the performance of the proposed dynamic consistent saliency detection model DCOF on the object detection task.

Based on the research of neuroscience, neurons in visual association cortex for example the inferior temporal cortex (IT), respond selectively to a particular object, especially to human faces. And the feedback originating in some higher level areas such as V4, V5, and IT can influence the human's attention in a top-down manner. From eye tracking experiments on image dataset, Judd et al. found that humans fixated so consistently on people and faces (Judd et al. 2009). Therefore, object detection result especially face detection region is often added into saliency map as a high level feature (Judd et al. 2009; Mathe & Sminchisescu 2012). Their experiment results prove that, as one kind of reliable high level information, the detection region of specific object detector is useful to build a better saliency map.

In this experiment, we compare our proposed dynamic consistent saliency detection model DCOF with two representative dynamic saliency models based on optical flow algorithms. The comparison models include the classical optical flow model (Horn & Schunck 1981; Black & Anandan 1996), and the spatial similarity optical flow model (Sun et al. 2010). First, we detect human's face in every video sample of Hollywood2 natural dynamic human scene videos dataset by the most commonly utilized face detector (Viola & Jones 2001). Then, we calculate the salient degree of the corresponding regions in the saliency detection results. The correct detection is defined as more than half of the pixels in the corresponding regions have larger saliency value than the average saliency value on the entire frame image.

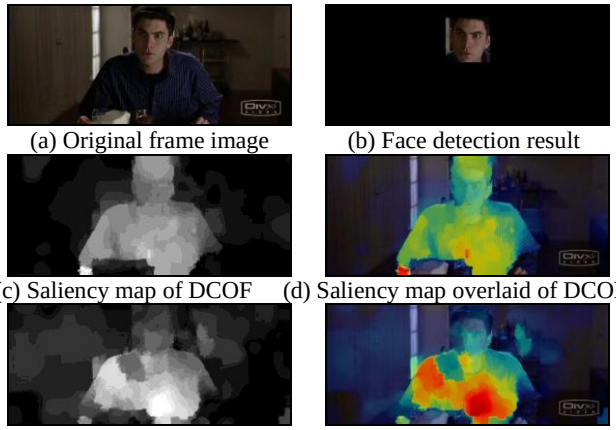
In Table 1, we list the average normalized saliency values of different models on face regions in the second column. The average detection accuracies of different models

are given in the third column. COF stands for classical optical flow model (Horn & Schunck 1981;Black & Anandan 1996), SOF stands for spatial similarity optical flow model (Sun et al. 2010). It is obvious that our model demonstrates best performance on both of evaluation standards.

Table 1. Face saliency detection on natural dynamic scene videos

Model	Average Saliency Value	Average Detection Accuracy
DCOF	**0.6501**	**0.8252**
COF	0.6018	0.7537
SOF	0.6393	0.7782

In Figure 1, we provide one example of temporal saliency detection results in this dataset. Figure 1(a) is an original frame image from 95th frame in autotrain00045 clip of scene videos labeled as kitchen. Figure 1(b) is the face detection results by (Viola & Jones, 2001). Figure (c) and (d) show the saliency map based on DCOF and the saliency map overlaid on the original image. Figure 1(e) and (f) provide the corresponding results of the spatial similarity optical flow model. Because the proposed dynamic saliency model emphasizes the dynamic continuity of neighbor locations in the same frame and same locations between the neighbor frames, the salient regions detected by our model can cover most of the informative areas of the image.



(a) Original frame image (b) Face detection result

(c) Saliency map of DCOF (d) Saliency map overlaid of DCOF

(e) Saliency map of SOF (f) Saliency map overlaid of SOF

Figure 1: Temporal saliency detection result.

The experiments on natural dynamic scene video reveal the motion information in video sequence often includes the prominent objects. Owing to the dynamic continuity and similarity is emphasized in our motion detection model, the extracted salient regions are often consistent with the object detection results. These results enlighten the proposed DCOF can be used to substitute the role of high-level features such as object detection feature maps in video saliency detection task.

Experiments on News Headline Videos

In the second experiment, we evaluate the efficiency of our proposed dynamic consistent saliency detection model DCOF on three typical CNN Headline news videos. Each of the video clips is approximately 30 seconds and the frame rate is 30 frames per second. The resolution of the frame image is 640×360. In this experiment, we also compare our model with two representative temporal saliency models based on optical flow algorithms, the classical optical flow model (Horn & Schunck 1981; Black & Anandan 1996), and the spatial similarity optical flow model (Sun et al. 2010).

We record the average running time per frame and the average output frame ratio over all frames in Table 2. The output frame ratio is defined as the number of the output motion frames divided by the number of video frames. All the codes are implemented in MATLAB R2012b on the test PC with Intel core I7-3520 2.9GHz and 4.00GB RAM. Because our model automatically determines the motion saliency group, the dynamic saliency map needn't to be calculated frame by frame. Therefore, our model demonstrates better efficiency and smaller storage capacity than existing models.

In Figure 2, one example of the dynamic saliency detection results of proposed model and the spatial similarity optical flow model is given. In the video sequence, the action of President Obama is not obvious between each adjacent frame pair. Unfortunately, the existing temporal saliency models based on optical flow method still estimate the motion between each adjacent frame pair independently. According to consider the dynamic consistency of neighbor locations in the same frame and same locations in temporal domain, our model can group the similar continuous action together. It reduces the running time and decreases the storage resource requirement. Furthermore, similar with the previous experiment results on natural scene videos, our dynamic consistent saliency detection model DCOF has a better coverage of the face than SOF.

Table 2. Efficiency comparison on the news headline videos

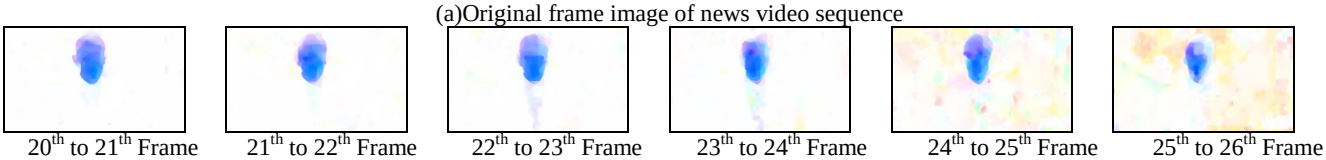
Model	Running Time (s)	Output Frame Ratio
DCOF	**33.12**	**0.4**
COF	46.24	1
SOF	53.88	1

Experiments on Eye-Tracking Action Videos

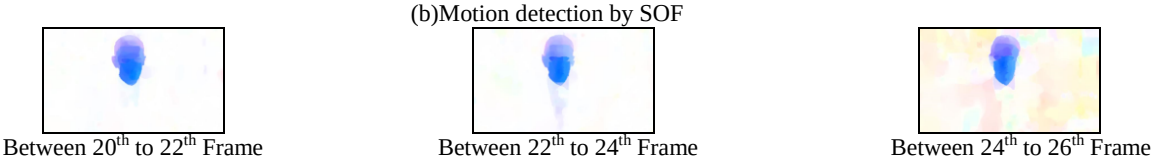
In the third experiment, we test our saliency detection model STA on the largest real world actions video dataset with human fixations (Mathe & Sminchisescu 2012). This action dataset contains 12 classes from 69 movies: answering phone, driving car, eating, fighting, getting out of car, shaking hands, hugging, kissing, running, sitting down, sitting up and standing up. The eye tracking data is collected by 16 subjects with low calibration error. In our experiment, we test on the first five videos from each category.

To evaluate the performance of various saliency models, we provide the results of the average receiver operating characteristic (ROC) curves and ROC areas. The ROC curve is plotted as the False Positive Rate vs. Hit Rate. The

(a)Original frame image of news video sequence



(b)Motion detection by SOF



(c)Motion detection by DCOF

Figure 2: Dynamic saliency detection result.

ROC area can be calculated as the area under the ROC curve to demonstrate the overall performance of a saliency model. Perfect prediction corresponds to the ROC area of 1.

Our model firstly compares with representative dynamic saliency models based on optical flow techniques (Horn & Schunck 1981; Black & Anandan 1996; Sun et al. 2010). The ROC area results are provided in Table 3. We could easily observe from Table 3 that DCOF has the largest ROC area and achieves the best overall performance.

Table 3. ROC area comparison on the eye-tracking action videos

ROC Area	DCOF	COF	SOF
Answer phone	**0.6098**	0.5303	0.5910
Drive car	**0.5233**	0.4817	0.5195
Eat	**0.6902**	0.6598	0.6644
Fight	**0.6045**	0.5535	0.6005
Get out car	**0.5260**	0.4874	0.5212
Hand shake	**0.6993**	0.6485	0.6934
Hug	**0.6402**	0.5602	0.5996
Kiss	**0.5833**	0.5120	0.5503
Run	**0.5535**	0.5104	0.5496
Sit down	**0.5183**	0.4761	0.5074
Sit up	**0.5171**	0.4871	0.5006
Stand up	**0.5602**	0.5269	0.5601

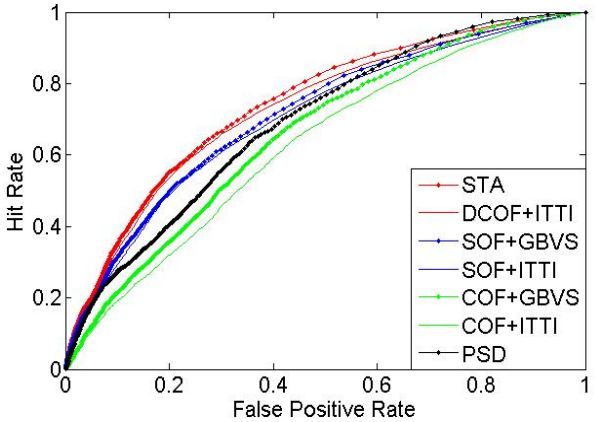


Figure 3: ROC curve comparison on eye-tracking videos.

Then, we combine some representative spatial saliency detection models with different dynamic saliency detection models. The spatial saliency models include: Itti saliency map (Itti et al. 1998) and graph based saliency map (Harel et al. 2007). In addition, the performance of the video saliency model predictive saliency detection model (PSD) is also provided here (Li et al. 2012). The average ROC curves comparison is shown in Figure 3. It can be seen that although all temporal saliency maps are benefited from integrating with the spatial saliency maps, STA model reaches the highest Hit Rate when False Positive Rate is low.

Conclusion

This paper proposes a novel spatial-temporal saliency detection model for video saliency detection. In spatial saliency map, we utilize the classical bottom-up spatial saliency features. In temporal saliency map, a novel optical flow model is proposed based on the dynamic consistency of motion. Two major advantages of the proposed model can be obtained: (1) Effective prominent object detection and coverage; and (2) Better efficiency and limited storage capacity. According to the empirical validation on three video datasets, the results show the performance of proposed dynamic consistent saliency model DCOF goes beyond the representative optical flow models and the state-of-the-art attention models. Experiment results also clearly demonstrate that the extracted salient regions by the proposed spatial-temporal attention model are consistent with the eye tracking data. Future work will be explored from two aspects. First, we will investigate how to explore our model to other real world applications. The second direction is to propose novel video attention model to jointly optimize the spatial and temporal saliency detection together.

Acknowledgments

This research was supported by HK PolyU 5245/09E.

References

Anderson, John R., 2004. Cognitive psychology and its implications. 6th Edition, Worth Publishers.

Avidan, S. & Shamir, A.. 2007. Seam carving for content-aware image resizing. In *ACM Transactions on Graphics*, pages 1-10.

Black, M. J. & Anandan, P. 1996. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. In *Computer Vision and Image Understanding*, 63, pages 75–104.

Born, R.T., Groh, J., Zhao, R., and Lukasewycz, S.J.. 2000. Segregation of object and background motion in visual Area MT: effects of microstimulation on eye movements, In *Neuron*, 26, pages 725-734.

Carrasco, M. & Frieder, K. 1997. Cortical magnification neutralizes the eccentricity effect in visual search. In *Vision Research*, 37(1), pages 63–82.

Cheng, W-H., Chu, W-T., Kuo, J-H., and Wu, J-L. 2005. Automatic video region-of-interest determination based on user attention model. In Proceedings of the IEEE International Symposium on Circuits and Systems. (ISCAS), pages 3219-3222.

Gibson, J.J.. 1950. The Perception of the Visual World. Houghton Mifflin.

Harel, J., Koch, C., and Perona, P.. 2007. Graph-based visual saliency. In Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS), pages 545-552.

Horn, B. & Schunck, B. 1981. Determining optical flow. In *Artificial Intelligence*, 16, pages 185–203.

Itti, L., Koch, C. and Niebur, E.. 1998. A model of saliency-based visual attention for rapid scene analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), pages 1254-1259.

Itti, L. & Baldi, P.. 2009. Bayesian surprise attracts human attention. In *Vision Research*, 49, pages 1295-1306.

Jiang, J.R. & Crookes, D.. 2012. Visual saliency estimation through manifold learning. In Proceedings of the 26th AAAI Conference on Artificial Intelligence, pages 2003-2010.

Judd, T., Ehinger, K., Durand, F. and Torralba, A.. 2009. Learning to predict where humans look. In Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV), pages 2106-2113.

Koch, C. & Ullman, S.. 1985. Shifts in selective visual attention: Towards the Underlying Neural Circuitry. In *Human Neurobiology*, pages 219-227.

Li, B., Xiong, W.H., Hu, W.M.. 2012. Visual saliency map from tensor analysis. In Proceedings of the 26th AAAI Conference on Artificial Intelligence, pages 1585-1591.

Li, Q., Chen, S.F., Zhang, B.W.. 2012. Predictive video saliency detection. In *Communications in Computer and Information Science*, 321, pages 178-185.

Li, Y. & Osher, S.. 2009. A new median formula with applications to PDE based denoising. In *Communications in Mathematical Sciences*, 7(3), pages 741–753.

Liu, C., Yuen, P.C., Qiu, G.P.. 2009. Object motion detection using information theoretic spatio-temporal saliency In *Pattern Recognition*, 42 (11), pages 2897–2906

Loy, C.C., Xiang, T., Gong S.G.. 2012. Salient motion detection in crowded scenes, In Proceedings of the 5th International Symposium on Communications, Control and Signal Processing (ISCCSP).

Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., Guérin-Dugué A.. 2009. Modelling spatio-temporal saliency to predict gaze direction for short videos. In *International Journal of Computer Vision*, 82(3), pages 231–243.

Marszałłek, M., Laptev, I., and Schmid, C.. 2009. Actions in Context. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

Mathe, S. & Sminchisescu, C.. 2012. Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition, In Proceedings of the 9th European Conference on Computer Vision (ECCV), pages 842-856.

Oikonomopoulos, I., Patras, I., Pantic, M.. 2006. Spatiotemporal salient points for visual recognition of human actions. In *IEEE Transactions on Systems, Man, and Cybernetics – Part B*, 36 (3), pages 710–719.

Parkhurst, D., Law, K., and Niebur, E.. 2002. Modeling the role of salience in the allocation of overt visual attention. In *Vision Research*, pages 107-113.

Peters, R. J., & Itti, L. 2008. Applying computational tools to predict gaze direction in interactive visual environments. In *ACM Transactions on Applied Perception*, 5.

Sun, D., Roth, S., Black, M. J.. 2010. Secrets of Optical Flow Estimation and Their Principles. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

Viola, P. & Jones, M.. 2001. Robust real-time object detection. In *International Journal of Computer Vision*.

Wang, T., Mei, T., Hua, X.-S., Liu, X., and Zhou, H.-Q. 2007. Video collage: A novel presentation of video sequence. In Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME), pages 1479--1482.

Vul, E., Frank, M.C., Tenenbaum, J.B., Alvarez, G.. 2010. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS), 22, pages 1955-1963.

Warren, D.H. & Strelow, E.R.. 1985. Electronic spatial sensing for the blind: contributions from perception. Martinus Nijhoff Publishers, Massachusetts.

Wedel, A., Pock, T., Zach, C., Cremers, D., and Bischof, H.. 2008. An improved algorithm for TV-L1 optical flow. In *Dagstuhl Motion Workshop*, 2008.

Xu, L., Jia,J. and Matsushita, Y.. 2010. Motion detail preserving optical flow estimation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

Yilmaz, A., Javed, O., and Shah, M.. 2006. Object Tracking: A Survey, In *ACM Computing Surveys*, 38(4), pages 13.

Yu, H., Li, J., Tian, Y., Huang, T., 2010. Automatic interesting object extraction from images using complementary saliency maps, In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM), pages 891-894.

Zhai, Y. & Shah, M.. 2006. Visual attention detection in video sequences using spatiotemporal cues. In Proceedings of the 14th ACM International Conference on Multimedia (ACMMM), pages 815-824.