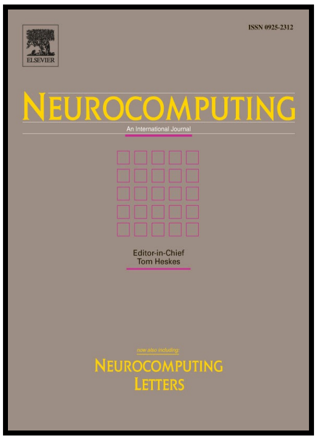
Author's Accepted Manuscript



www.elsevier.com/locate/neucom

Perception-oriented Video Saliency Detection via Spatio-Temporal Attention Analysis

Sheng-hua Zhong, Yan Liu, To-Yee Ng, Yang Liu

PII:	S0925-2312(16)30327-7
DOI:	http://dx.doi.org/10.1016/j.neucom.2016.04.048
Reference:	NEUCOM17018
To appear in:	*Neurocomputing*
Received date:	7 December 2015
Revised date:	7 March 2016
Accepted date:	27 April 2016

Cite this article as: Sheng-hua Zhong, Yan Liu, To-Yee Ng and Yang Liu, Perception-oriented Video Saliency Detection via Spatio-Temporal Attentior Analysis, *Neurocomputing,* http://dx.doi.org/10.1016/j.neucom.2016.04.048

Perception-oriented Video Saliency Detection via Spatio-Temporal Attention Analysis

Sheng-hua Zhong[a, b], Yan Liu[c,*], To-Yee Ng[c], Yang Liu[d]

[a] College of Computer Science and Software Engineering, Shenzhen University, Shen Zhen, PR China

[b] Department of Psychological & Brain Sciences, The Johns Hopkins University, Baltimore, MD, U.S.

[c] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, PR China

[d] Department of Computer Science, The Hong Kong Baptist University, Hong Kong, PR China

Abstract

Human visual system actively seeks salient regions and movements in video sequences to reduce the search effort. Computational visual saliency detection model provides important information for semantic understanding in many real world applications. In this paper, we propose a novel perception-oriented video saliency detection model to detect the attended regions for both interesting objects and dominant motions in video sequences. Based on the visual orientation inhomogeneity of human perception, a novel spatial saliency detection technique called visual orientation inhomogeneous saliency model is proposed. In temporal saliency detection, a novel optical flow model is created based on the dynamic consistency of motion. We fused the spatial and the temporal saliency maps together to build the spatio-temporal attention analysis model toward a uniform framework. The proposed model is evaluated on three typical video datasets with six visual saliency detection algorithms and achieves remarkable performance. Empirical validations demonstrate the salient regions detected by the proposed model highlight the dominant and interesting objects effectively and efficiently. More importantly, the saliency regions detected by the proposed model are consistent with human subjective eye tracking data.

Keywords: Perception-oriented video saliency; spatio-temporal modeling; orientation inhomogeneous feature map; dynamic consistency; visual attention.

1. Introduction

Visual perception is an active process to interpret the surrounding environment by processing information contained in visual light [1]. Theories and observations of visual perception have been the main source of inspiration for computer vision and artificial intelligence. Perceptual models play an increasingly significant role in optimizations of various applications, such as perceptual-based video coding [2], quality assessment [3], real-time user authentication [4], sound classification [5], and audio watermarking [6].

Visual attention analysis plays an important role in perception-oriented modeling and attracts interest from a broad range of researchers and scientists. First, attention is the behavioral and cognitive process of selectively concentrating on one aspect of the environment while ignoring other things [7]. In human, attention is facilitated by a retina that has evolved a high-resolution central fovea and a low-resolution periphery [8], which is also closely related to multiple other cognitive processes, such as: perception, memory, and learning. Hence, the research on visual attention analysis provides a window on the human perception and other cognitive processes. Second, based on the eye tracking data recorded by the high-speed eye tracker, attention is more easily to be observed than other cognitive processes. Third, the research on attention analysis provides an effective means to connect human perception and the various applications in further processing [9], including image quality assessment [10], object detection [11], action recognition [12], image retargeting [13], video abstraction [14], removing label ambiguity in image [15], etc.

The strongest attractors of attention are stimuli that pop-out from their neighbors in space or time usually referred to as “saliency” [16]. Visual attention analysis simulates the human visual system behavior by automatically producing saliency maps of the target image or video sequence [17]. The saliency map is proposed to measure the conspicuity and calculate the likelihood of a location in visual data to attract attention [18]. Therefore, the visual saliency detection provides predictions on which regions are likely to attract observers’ attention [19]. Although image saliency detection has been long studied, little work has been extended to video sequences due to the data complexity. After the standard real world video datasets with subjects’ eye tracking data emerge, such as the video action dataset [20], a more detailed and quantitative research for video saliency detection and analysis will be feasible.

The conference version of our preliminary work was published in [21]. This work demonstrates good performance on saliency detection based on dynamic consistency of motion. But in spatial saliency modelling, it inherits the classical bottom-up spatial saliency map. In this paper, we propose a novel perception-oriented video saliency detection method called spatio-temporal attention analysis model (STAM) by referring to the characters of the human visual system. The STAM follows the three-part scheme of video saliency detection, including spatial saliency detection, temporal saliency detection, and the fusion of spatial and temporal saliency maps. In feature extraction stage of spatial saliency detection, multiple low-level visual features including: intensity, color, orientation, and contrast are extracted at multiple scales. Instead of using the original orientation feature map, we propose a novel technique called visual orientation inhomogeneous saliency model (VOIS). In our orientation feature map, the information in cardinal orientations is retained, but the information in oblique orientations is weakened with the inhomogenous weight. Then, the activation maps are built based on multiple low-level feature maps. And the saliency map is finally constructed by a normalized combination of the activation map. In temporal saliency map modeling part, a novel dynamic consistent optical flow model (DCOF) is proposed based on the human visual dynamic continuity. Different from the classical optical flow model which estimates motion between each adjacent frame pair independently, the proposed DCOF takes account of the motion consistency in video sequence. In saliency fusion stage, the “skew-max” fusion method is utilized to fuse the spatial and temporal saliency maps together and construct the final video saliency map.

In the following parts of this paper, we discuss the related work on video saliency detection in Section 2. A novel spatio-temporal video saliency detection technique is introduced in Section 3. In Section 4, we demonstrate the performance of the proposed video saliency detection model on three video sequence datasets. The paper is closed with conclusion and future work in Section 5.

2. Related work on Saliency detection

Video saliency detection calculates the salient degree of each location by comparing with its neighbors both in spatial and in temporal areas. Previously, most existing computational saliency models depend on the intrinsic bottom-up spatial features of the visual stimuli by referring to the human visual system [22] [23]. Neurophysiological experiments have proved that neurons in the middle temporal visual area (MT) compute local motion contrast. Such neurons, which underlie the perception of motion pop-out and figure-ground segmentation, influence the attention allocation [24]. After realizing the importance of motion information in video attention, the motion feature has been added into the saliency models [25][26]. Recently, to simulate two pathways (parvocellular and magnocellular) of the human visual system, the video saliency detection procedure is divided into spatial and temporal pathways [27]. These two pathways (the P and M pathways) correspond to the static and dynamic information of video. In P pathway, parvocellular cell has greater spatial resolution, but lower temporal resolution. Conversely, in M pathway, magnocellular cell has greater temporal resolution, but lower spatial resolution.

Typically, in spatial pathway saliency detection, most of the video saliency techniques follow the classical image saliency architecture including three stages: feature extraction, activation, and normalization. Multiple low-level visual features such as intensity, color, orientation, and contrast are firstly extracted at multiple scales. Then, the activation maps are built based on multiple low-level feature maps. After the activation maps are computed, they are normalized and combined into a spatial saliency map that represents the saliency of each pixel [28]. Almost all of the existing bottom-up models are inspired by the theories from human visual system [29]. Among them, the most famous one was proposed by Itti *et al.* [30]. They developed the center surround structure akin to on-type and off-type visual receptive field. We denote this model as ITTI in our experiment. In recent years, more proposed work simulated the multi-scale and multi-orientation function of primary visual cortex. Achanta *et al.* detected the saliency map with a Difference of Gaussians (DOG) model to describe the spatial properties of visual regions [31]. Gabor filters and Log-Gabor wavelets were utilized to explore the salient features such as spatial localization, spatial frequency characteristics in [32] and [29], respectively.

In temporal pathway saliency detection, optical flow is the most widely used method in existing video saliency detection models [20] [31] [33]. These models rely on the classical optical flow method to extract the motion vector between each frame pair independently as the temporal saliency map. The classical formulation of optical flow was first introduced by Horn and Schunck [34]. They optimized 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 further addressed the outlier sensitivity problem of initial optical flow model by replacing the quadratic error function with a robust formulation [35]. Although different efforts have been put into improving the optical flow, the median filtering is the most important source to improve the performance of the classical optical flow model [36]. According to the extensive test by [37], the median filtering makes non-robust methods more robust and improves the accuracy of the optical flow models. Unfortunately, although the optical flow techniques can accurately detect the 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 should be popped out as the indicator of the temporal salient region.

3. Spatio-temporal attention model

In this section, we propose a novel spatio-temporal attention analysis model (STAM). The schematic illustration of the proposed STAM is described in Fig. 1.

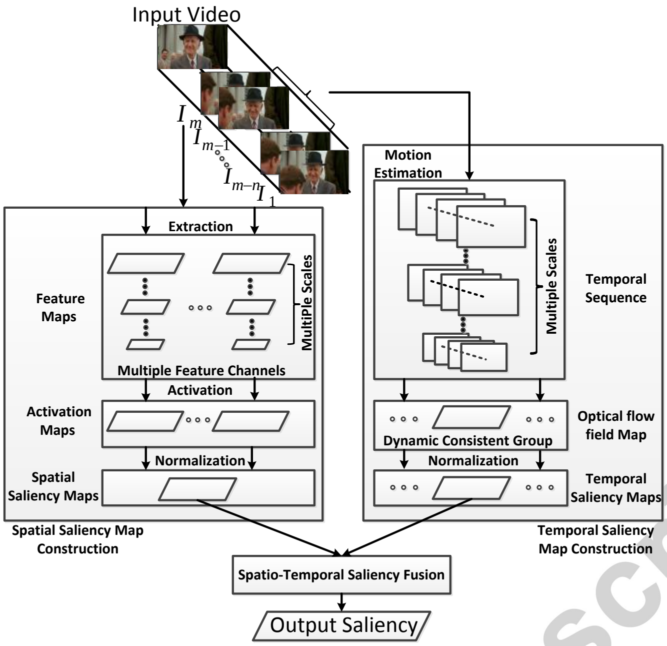


Fig. 1. Schematic illustration of spatio-temporal attention analysis model (STAM).

The whole spatio-temporal attention analysis model can be partitioned into two pathways. In spatial saliency map construction, we follow the three common stages of the classical bottom-up spatial saliency map. A novel spatial saliency map technique called visual orientation inhomogeneous saliency model (VOIS) is proposed in Section 3.1. In VOIS, we will provide a human-like orientation feature map extraction based on the visual orientation inhomogeneity of human perception. In temporal saliency map construction, a novel dynamic consistent optical flow model (DCOF) is proposed in Section 3.2 based on the human visual dynamic continuity. Different from the classical optical flow model estimates motion between each adjacent frame pair independently, DCOF both underlines the consistency of motion saliency in the current frame and between the consecutive frames. In Section 3.3, we simply adopt the “skew-max” fusion method from existing work to obtain the final video saliency map.

3.1. Spatial saliency map construction

The leading models of spatial saliency map construction can be divided into three stages:

1) Extraction: multiple low-level visual features (intensity, color, orientation, and contrast) 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 this part, we propose a novel spatial saliency detection technique called visual orientation inhomogeneous saliency model (VOIS) under this three-stage learning procedure. Specifically, based on the visual orientation inhomogeneity, in the first stage, we try to provide a human-like orientation feature map extraction to substitute the original orientation feature map.

In the first stage, the bottom-up features are extracted based on several feature channels. Among these feature channels, the orientation information is known to play an important role in early visual system. Based on the discovery of Hubel and Wiesel, a key characteristic of the responses of primary visual area (V1) neurons is their high selectivity for stimulus orientation [38]. Most early visual neurons tuned to some type of local spatial contrast (such as center-surround or orientated edges) [39]. The examinations of the early visual system of mammals have shown Gabor-like behavior of the simple cell responses [40]. The orientation information is thought to be a basic component of an object [41]. Girshick *et al.* [42] found the human observation exploits perception inhomogeneities in orientation. It means the human observation is worst at oblique angles and best at cardinal (horizontal and vertical) angles. From the physiological instantiation, they found that the non-uniformities in the representation of orientation in the V1 population contribute to non-uniformities in perceptual discriminability. Specifically, a variety of measurements have shown that cardinal orientation is represented by a disproportionately large fraction of V1 neurons, and that those neurons also tend to have narrower tuning curves [43]. Based on these proofs from neuroscience, Brecht *et al.* [44] kept the cardinal orientation feature map to implement a neural network for human visual search mechanism simulation. In Fig. 2, an example of orientation feature map extraction in cardinal and oblique orientations in multiple scales is demonstrated.

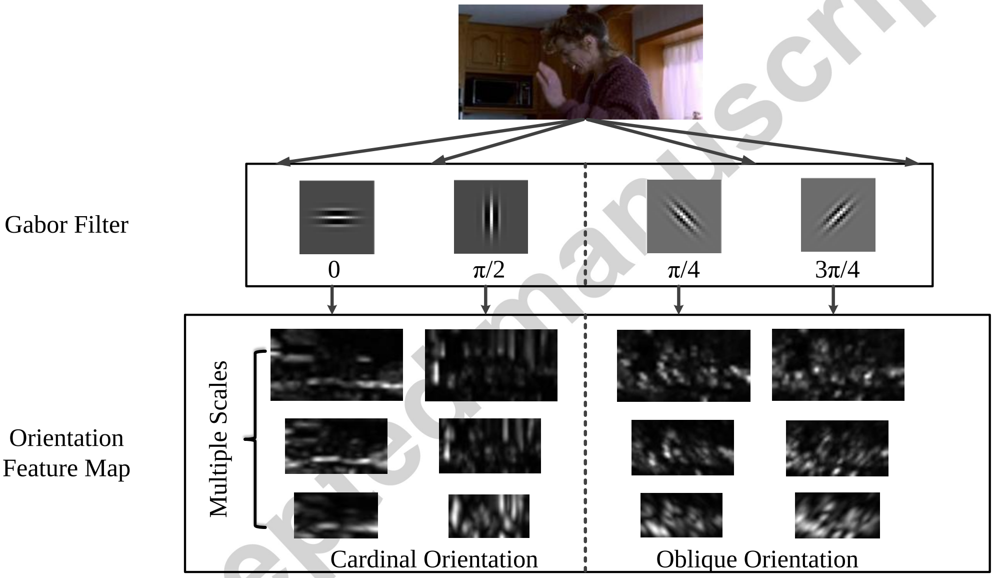


Fig. 2. Example of orientation feature map extraction in cardinal and oblique orientations in multiple scales.

In the proposed spatial saliency map construction, unlike the existing standard orientation feature maps, the novel orientation feature map extracts the non-uniformly information from different visual orientation as Eq. (1):

$$F^o = F_c + \alpha F_o \tag{1}$$

where F^o is the proposed feature map in orientation channel. F_c is the feature map extracted in cardinal orientations, and F_o is the feature map extracted in oblique orientations. The information in cardinal orientations is retained, but the information in oblique orientations is weakened with the inhomogenous weight parameter α ($0 \le \alpha \le 1$). If $\alpha = 1$, it is the original orientation map whose the cardinal and oblique orientations are treated equally. If $\alpha = 0$, it is a special version of orientation feature maps that just includes cardinal orientations. As we known, one of the underlying reasons for the anisotropy of orientation discriminability is the prevalence of vertical and horizontal orientations in the real-world [42]. By utilizing Principal Component Analysis (PCA) to obtain the principal components for energy spectra of real-world scenes, Oliva and Torralba found that vertical and horizontal orientations are more frequent than obliques [45]. They obtained more precise anisotropy in image category by fitting the distribution of orientations to the power spectrum [46]. Zhong *et al.* validated that visual orientation inhomogeneity descriptor can achieve better or at least comparable performance with less computation resource and time in various computer vision tasks under real world conditions, such as image matching and object recognition [47]. Similar with these existing researches, the statistical analysis of the orientation distribution can be used to infer the inhomogenous weight in oblique orientation of the orientation feature map extraction. Therefore, in this paper, the inhomogenous weight α is obtained by referring to the orientation distribution in environment.

To obtain the inhomogenous weight, we statistically analyze the orientation distribution in environment on the standard dataset Urban and Natural Scene dataset [45]. This dataset includes 2,688 authentic images with eight semantically organized categories, namely “coast,” “forest,” “highway,” “city center,” “mountain,” “open country,” “street,” and “tall building.” Here, we define the environmental orientation distribution as the probability distribution over local orientations with different spatial scale. The Canny edge detector [48] is applied to form the edge map of every authentic image in Urban and Natural Scene dataset. The threshold of the Canny detector is set according to the default setting of Matlab edge detection techniques. The local image gradients are computed based on the edge map. The orientation histograms are statistically calculated based on the gradient orientation values. In Table 1, we compare the orientation magnitude proportion of cardinal orientations with oblique orientations in eight different categories of the Urban and Natural Scene dataset. The statistical significance of the difference between the cardinal orientations *vs.* oblique orientations is tested on paired *t* tests. According to the results of *t* test, the difference between them is significant. We also tried Garbor filter as the tool to calculate the orientation magnitude proportion. We find these results are similar.

The inhomogenous weight α is simply defined as Eq. (2):

$$\alpha = \frac{\overline{P}_o}{\overline{P}_c} \tag{2}$$

where $\overline{P}_o$ is the average value of the cardinal orientations magnitude proportion. $\overline{P}_c$ is the average value of the oblique orientations magnitude proportion.

Table 1. Magnitude proportion of cardinal vs. oblique orientation

Category	Orientation	Mean±Sem	*P* Value
Coast	Cardinal	**0.6999±0.00406**	<0.0001
	Oblique	0.3001±0.00406	
Forest	Cardinal	**0.5399±0.00407**	<0.0001
	Oblique	0.4601±0.00407	
Highway	Cardinal	**0.6564±0.00532**	<0.0001
	Oblique	0.3436±0.00532	
City center	Cardinal	**0.7539±0.00487**	<0.0001
	Oblique	0.2461±0.00487	
Mountain	Cardinal	**0.5436±0.00228**	<0.0001
	Oblique	0.4564±0.00228	
Open country	Cardinal	**0.5718±0.00280**	<0.0001
	Oblique	0.4282±0.00280	
Street	Cardinal	**0.6255±0.00335**	<0.0001
	Oblique	0.3745±0.00335	
Tall building	Cardinal	**0.7027±0.00505**	<0.0001
	Oblique	0.2973±0.00505	

To the activation and normalization stage, we follow the technique of graph-based saliency (GBVS) [32] to construct the spatial saliency map. Similar with GBVS, in our method, the fully-connected directed graph G_A is constructed in activation stage. The weight of the directed edge in G_A from node (i, j) to node (p, q) is assigned as Eq. (3):

$$w_A((i,j),(p,q)) = d((i,j)\,\|\,(p,q)) \cdot G(i-p, j-q) \tag{3}$$

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

(4)

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.

In normalization stage of our method, the fully-connected directed graph G_N is built. The weight of the directed edge in G_N from each node (i,j) to node (p,q) is assigned as Eq. (5), where A is the activation map.

$$w_N((i,j),(p,q)) = A(p,q) \cdot G(i-p, j-q) \tag{5}$$

Then, we follow the technique of graph-based saliency. The spatial saliency map **SSMap** is formed based on the fully-connected directed graph G_A and G_N.

3.2. Temporal saliency map construction

As we described before, in the existing video saliency detection models, optical flow technique is the most widely used temporal saliency detection approach [20] [31] [33]. The optical flow 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 [49]. It was first studied in the 1940s and ultimately published by psychologist [50]. The optical flow approach approximates the object motion by estimating vectors originating or terminating at pixels in image sequences, so it represents the velocity field that warps one image into another feature space [51]. The motion detection methods based on optical flow technique can accurately detect the motion in the direction of intensity gradient.

The objective function of the classical optical flow is defined as:

$$\begin{aligned} E(\mathbf{u},\mathbf{v}) = \sum_{i,j} \{ & f_D(I_1(i,j) - I_2(i+u_{i,j}, j+v_{i,j})) \\ & + \lambda_1 [f_S(u_{i,j} - u_{i+1,j}) + f_S(u_{i,j} - u_{i,j+1}) + f_S(v_{i,j} - v_{i+1,j}) + f_S(v_{i,j} - v_{i,j+1})] \} \end{aligned} \tag{6}$$

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_1 is a regularization parameter. f_D is the brightness constancy penalty function, and f_S is the smooth penalty function. Here, we refer to the formulation in Eq. (6) and all the formulations that are directly derived from it as the “classical optical flow model”.

Most of the existing temporal saliency detection techniques rely on the classical optical flow method to extract the motion vector between each frame pair independently. Although the optical flow technique can detect motion accurately, the temporal saliency is not perfectly equal to the amplitude of all the motions between each adjacent frame pair. In fact, some subtle motions between frames are often resulted from the illumination change or other unsteady small-disturbance in environment. Therefore, in this case, the motion of the prominent object is possible to be drowned in optical flow vectors. Furthermore, 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):

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

where I_m is the m^{th} frame in video sequence $\mathbf{X}$, n is the number of neighbor frames with consistent motion. Here, $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ denote two components of the auxiliary flow field. λ_1, λ_2 and λ_3 are the scalar regularization weights. The alternating optimization strategy is used to compute the objective function in Eq. (7). The first term of Eq. (7) emphasizes the dynamic continuity of same locations in temporal domain. The second term the dynamic continuity of neighbor locations in same frame. The third term encourages the auxiliary flow field $\hat{\mathbf{u}}$, $\hat{\mathbf{v}}$ and the flow field $\mathbf{u}$, $\mathbf{v}$ to be the same. The last term of the optimal function imposes a smoothness assumption within a region corresponding to the auxiliary flow field. The optimal function is constrained in a reasonable range by imposing the constraint based on the observation standard deviation $\sigma_o(i,j)$ of human visual perception.

Given by the eccentricity scaling parameter *c*, $\sigma_o(i,j)$ is calculated by Eq. (8) [52][53], where γ is denoted as the number of pixels per degree of visual angle, γ is the first-order coefficient which is equal to 0.42 [52].

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

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.

Algorithm 1: Dynamic Consistent Saliency Detection

```
Input:   Video sequence data $\mathbf{X}$;
         Number of frames $N_f$;   $m$=1; $n_0$=5; Pyramid level $N_p$.
Output: Temporal saliency map TSmap.

while $m < N_f$ - $n$
    $p$ = 1; $n$ = $n_0$;
    while $p < N_p + 1$ do
        $(n,\mathbf{u},\mathbf{v})=\arg\min_{n,\mathbf{u},\mathbf{v}} E(\mathbf{u},\mathbf{v},\hat{\mathbf{u}},\hat{\mathbf{v}})$;
        $(i^*,j^*)=\arg\max_{i,j}(\sqrt{{u_{i,j}}^2+{v_{i,j}}^2})$;
        if $\sqrt{{u_{i^*,j^*}}^2+{v_{i^*,j^*}}^2}>\frac{\sigma_o(i^*,j^*)}{2^{p-1}}$ and $n$ > 1
            $n$ = max($n$ - 1,1); $p$ = 1;
        else
            $p$ = $p$ + 1;
        end if
    end while
    TSmap$_m(i,j)$ = normalize($\sqrt{{u_{i,j}}^2+{v_{i,j}}^2}$);
    $m$ = $m$ + $n$;
end while
```

3.3. Spatio-Temporal saliency Fusion

In this stage, we fuse the spatial and temporal saliency maps together to get the final video saliency map. As we known, different fusion methods have been proposed and utilized, such as "mean" fusion, "max" fusion, "multiplicative" fusion, and "skew-max" fusion.

The "mean" fusion takes the pixel average of the spatial and temporal saliency maps:

$$\mathbf{FSmap} = (\mathbf{SSmap}+\mathbf{TSmap})/2 \tag{9}$$

The "max" fusion constructs each pixel as the maximum of the two saliency maps:

$$\mathbf{FSmap} = \max(\mathbf{SSmap},\mathbf{TSmap}) \tag{10}$$

A pixel by pixel multiplication operation is the "multiplicative" fusion:

$$\mathbf{FSmap = SSmap}\times\mathbf{TSmap} \tag{11}$$

"Skew-max" fusion both takes advantage of the characteristics of the spatial and the temporal saliency maps:

$$\begin{aligned}\mathbf{FSmap} =& \max(\mathbf{SSmap})\times\mathbf{SSmap}+\text{skewness}(\mathbf{TSmap})\times\mathbf{TSmap} + \max(\mathbf{SSmap}) \\ &\times\text{skewness}(\mathbf{TSmap})\times\mathbf{SSmap}\times\mathbf{TSmap}\end{aligned} \tag{12}$$

The "mean" fusion modulates one map with the other. If a pixel is salient for the temporal map but not for the spatial one, the fusion saliency result is lower than it in the spatial one. For the "max" fusion, a pixel has the highest saliency between the spatial and temporal maps and is less selective. The "multiplicative" fusion is the most selective one. In previous work, the saliency maps of "mean" and "max" fusions demonstrate close performance [54]. Marat *et al.* indicated that "skew-max" integration method [55] has best performance as the spatial-temporal integration function [27]. Hence, in this paper, we simply adopt the "skew-max" fusion method to take for the combination of two saliency maps as Eq. (12).

4. Empirical validation

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 saliency detection model is tested on the Hollywood-2 natural dynamic human scene videos dataset [55]. In this dataset, we want to demonstrate the performance of the proposed saliency detection model and other saliency detection models for the object detection task. In the second experiment, three typical News videos collected from YouTube are utilized to test the efficiency of the proposed model. The third experiment is evaluated on the largest real world actions video dataset with human fixations [20]. In this dataset, the salient degree is measured in accordance with attention allocations of human based on the fixations of subjects.

In spatial saliency detection, most of techniques include an independent stage to extract low-level visual features including orientation. Among these models, two classical bottom-up models are effective and accurate for detecting salient regions. Our model is compared with these two classical spatial models [30][32] to explore the effect of the orientation inhomogeneity. As we known, optical flow technique is the most widely used temporal saliency detection approach. And in temporal saliency detection, we proposed a novel optical flow model based on the dynamic consistency of motion. Thus, our comparison also includes two representative optical flow models [34][35][37]. In model comparison, we also combined the spatial saliency map model [30] and [32] with these dynamic saliency detection models. Moreover, we compared with two state of the art visual saliency detection methods, self-resemblance [56] and signature saliency model [57].

For the functions and parameters in the proposed DCOF, we simply adopt the general setting of optical flow model in existing papers as follows. For example, Convex Charbonnier penalty function is implemented as the penalty function f_D and f_S. The number of warping steps per pyramid level is set as 3. The regularization parameter λ_1 is selected to be 5. We perform ten steps of alternating optimization at every pyramid level and change λ_2 logarithmically from 10^{-4} to 10^2. The scalar weight λ_3 is set as 1. A 5×5 size rectangular window is set as the smooth region of the flow field. To the eccentricity scaling parameter *c*, we just follow the general setting in [53] to set *c* as 0.08. To the parameter σ_G, we simply follow the setting of GBVS in [32]. The parameter σ_G is set as 5. To the parameters in self-resemblance algorithm [56] and signature saliency model [57], we follow all standard setting in the release Matlab Toolbox from the authors.

4.1. Experiments on Natural Human Scene Videos

In the first experiment, we evaluate the proposed saliency detection models on the Hollywood-2 natural human scene videos dataset [55]. This dataset contains the natural dynamic samples of human in ten different natural categories, including: “house”, “road”, “bedroom”, “car”, “hotel”, “kitchen”, “living room”, “office”, “restaurant”, and “shop”. It consists of about ten hours of Hollywood movies and is split into 1152 video sequences. In this experiment, we want to demonstrate the performance of the proposed spatial and temporal saliency detection models for the object detection task.

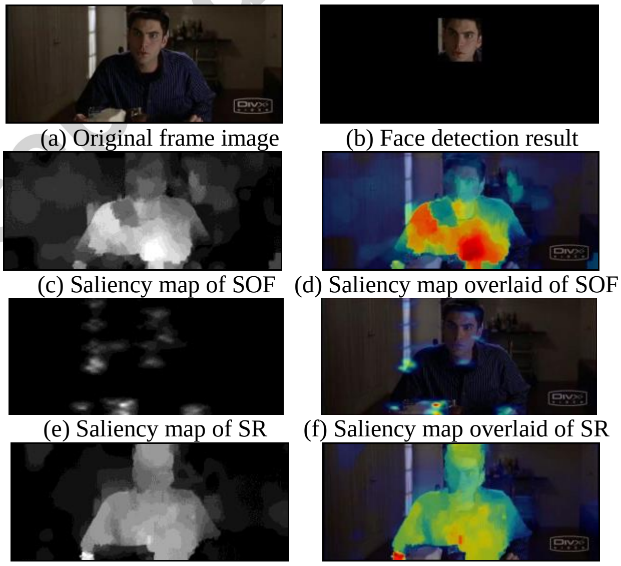
Based on the research of neuroscience, neurons in visual association cortex for example the inferior temporal cortex (IT), respond selectively to a particular object, especially to human faces. And the feedback originating in some higher level areas such as V4, V5, and IT can influence the human’s attention 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 [58]. Therefore, the object detection result especially the face detection region is often added into the final saliency map as a high level feature [20][58]. The objectness likelihood is also integrated into computational algorithm [59] to detect visual saliency. Their experimental results also evidence that, as one kind of reliable high-level information, the detection region of specific object detector is useful to build a better saliency map. Meanwhile, as we known, the good low-level saliency map should work as indicators for the object detection and recognition, especially the temporal saliency map based on motion detection. In our proposed temporal saliency detection model, we also assume the dynamic continuity could help us get the motion of the prominent object. Hence, in this experiment, we first compare the proposed VOIS with the representative spatial saliency models graph based saliency map [32]. Then, the proposed DCOF is compared with three saliency detection models, including two representative temporal saliency models [34][35][37] and one state of the art spatial-temporal saliency model [56]. First, we detect human’s face in every video sample of Hollywood-2 natural dynamic human scene videos dataset by the most commonly utilized face detector [60]. Then, all of the saliency detection models are used to generate the saliency maps. Third, the salient degree of the corresponding face regions in the saliency maps is calculated.

In Table 2, we provide the average normalized salient values of different models on face regions in the second column. Moreover, the average detection accuracies of different models are given in the third column. Here, the correct detection is defined as more than half of the pixels in the corresponding regions have larger salient value than the threshold value 0.5 on the entire frame image. COF stands for classical optical flow model [34][35], SOF stands for spatial similarity optical flow model [37]. SR stands for the self-resemblance model based on spatial and temporal channels [56]. It is obvious that our model demonstrates best performance on both of evaluation standards and also both of pathways. From Table 2, we can find the temporal saliency map techniques achieve better performance in face saliency detection than the spatial saliency map techniques. This is because, in the real-world video sequences, such as movies, the human face region is often the part containing apparent motion.

Table 2. Face Saliency detection of different saliency model on Natural Dynamic Scene Videos

Model	Average Salient Value	Average Detection Accuracy
VOIS	**0.5281**	**0.7653**
GBVS	0.4589	0.7375
COF	0.6018	0.7537
SOF	0.6393	0.7782
SR	0.4624	0.7221
DCOF	**0.6501**	**0.8252**

In Fig. 3, we provide one example of the saliency detection results in this dataset. Fig. 3(a) is an original frame image from 95th frame in autotrain00045 clip of scene videos labeled as kitchen. Fig. 3(b) is the face detection results by [60]. Figs. 3(c) and 3(d) show the saliency map based on the spatial similarity optical flow model SOF and the saliency map overlaid on the original image. Figs. 3(g) and 3(h) provide the corresponding results of the proposed model DCOF. Here, we also provide the results in Fig. 3(e) and 3(f) based on self-resemblance [56], a video saliency detection model with spatial-temporal channels. According to the comparison, we can find 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. The experimental results on natural dynamic scene video are even better than the self-resemblance model based on spatial and temporal channels [56]. These results also reveal the motion information in video sequence often includes the prominent objects, such as human faces. It enlightens that 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.



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

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

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

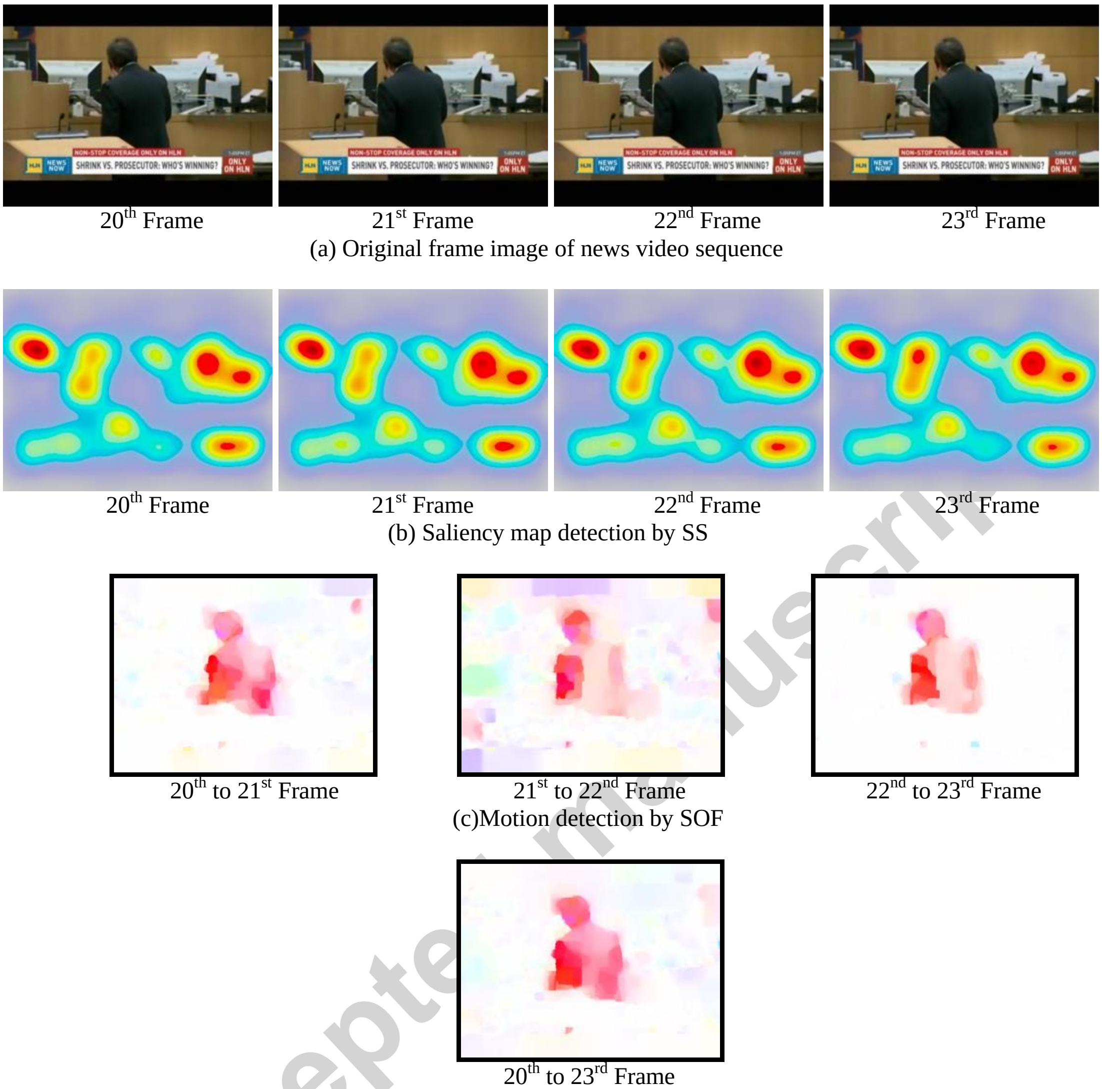
(g) Saliency map of DCOF (h) Saliency map overlaid of DCOF

Fig. 3. Temporal saliency detection results.

4.2. Experiments on News Headline Videos

As one kind of typical video sequences, the news headline videos are often selected as the materials to analyze the video saliency detection models, especially the temporal saliency detection models [61]. In the existing temporal saliency detection techniques, the widely used optical flow approaches rely on the motion vector between each frame pair independently. Unfortunately, although the optical flow techniques can accurately detect the motion in the direction of intensity gradient, it has some limitations in temporal saliency detection. The temporal saliency is not perfectly equal to the amplitude of all the motion between each frame pair. If some changes between adjacent frame pair exist in video, such as the change of illumination, or other kind of noise, it can be judged as the temporal salience by these optical flow techniques. In this case, the motion of the prominent object is possible to fade into the background. In addition, the independent calculation of each frame pair brings about a high computational complexity. Hence, to evaluate the efficiency performance of the proposed DCOF, in the second experiment, we test 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 480×360. In this experiment, we compare our model with two representative temporal saliency models based on optical flow algorithms: the classical optical flow model [34][35] and the spatial similarity optical flow model [37].

We record the average running time per frame and the average output frame ratio over all frames in Table 3. 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.



20th Frame 21st Frame 22nd Frame 23rd Frame

(a) Original frame image of news video sequence

20th Frame 21st Frame 22nd Frame 23rd Frame

(b) Saliency map detection by SS

20th to 21st Frame 21st to 22nd Frame 22nd to 23rd Frame

(c)Motion detection by SOF

20th to 23rd Frame

(d)Motion detection by DCOF

Fig. 4. Dynamic saliency detection result.

In Fig. 4, an example of the dynamic saliency detection results comparison is given. Fig. 4(a) demonstrates the original frame image of news video sequence. Fig. 4(b) is the saliency detection result of signature saliency model (SS) based on spatial pathway [57]. From these results, it is obvious that the spatial saliency map, which only catches the parts with high contrast, does not match the dynamic saliency in the video. In the video sequence, the action of person 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 as Fig. 4(c). According to considering the dynamic consistency of neighbor locations in the same frame and same locations in temporal domain, our temporal model can group the similar continuous actions together. Therefore, it reduces the running time and decreases the storage resource requirement. Furthermore, there exists some change of illumination or other noises from 21st frame to 22nd frame. Although the change is not obvious in original video frames, it still influences the motion detection by SOF. The motion of some part in the person's body is even drowned in optical flow vectors of SOF. Our proposed DCOF achieves better coverage of the prominent objects in Fig. 4(d).

Table 3. 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

4.3. Experiments on Eye-Tracking Action Videos

Recently, tremendous attempts have been made in using perceptual saliency models to calculate the salient degree in accordance with human subjective attention allocation based on the eye tracking data. Therefore, we also test the proposed STAM on the largest real world actions video dataset with human fixations [20]. This action dataset is one of the most challenging available for real world actions. It contains 12 classes from 69 movies: "AnswerPhone," "DriveCar," "Eat," "FightPerson," "GetOutCar," "HandShake," "HugPerson," "Kiss," "Run," "SitDown," "SitUp" and "StandUp." The eye moments were recorded using an SMI iView X HiSpeed 1250 tower-mounted eye tracker, with a sampling frequency of 500 HZ. The head of the subject was placed on a chin-rest located at 60cm from the display. The LCD display had a resolution of 1280×1024 pixels, with a physical screen size of 47.5×29.5cm. The tracking data is collected from 16 human volunteers (9 males and 7 females) aged between 21 and 41 with low calibration error. In our experiment, we test on the first five videos of each category.

To evaluate the performance of various saliency models, we provide the results of the average receiver operating characteristic (ROC) areas. The ROC curve is created by the fraction of true positive out of the positives (TPR = true positive rate) vs. the fraction of false positive out of the negatives (FPR = false positive rate), at various threshold settings. The ROC area can be calculated as the area under the ROC curve, and it demonstrates the overall performance of a saliency model.

To evaluate the performance of our model, we provide the comparison of the ROC area coverage in spatial saliency map and temporal saliency map, respectively. In Table 4, we compare the proposed VOIS with the existing spatial saliency models, including: graph based saliency map [32], and the CAL_VOIS. CAL_VOIS is a simple version of VOIS whose orientation feature map just includes cardinal orientations. From Table 4, obviously, the proposed VOIS can cover larger ROC area than other existing spatial saliency map based on the standard orientation feature map in each category. Furthermore, even the information in oblique orientations are ignored, the ROC area coverage is comparable to GBVS. The statistical significance of the difference between the VOIS vs. GBVS is tested on paired *t* tests. According to the results of *t* test, the difference between them is significant (p<0.001). Then, we compare the area coverage of the proposed temporal saliency map and representative temporal saliency models based on previous optical flow techniques [34][35][37]. The ROC area results are provided in Table 4. We could easily observe the proposed DCOF achieves the largest coverage of the fixation points. In most of categories, the temporal saliency models achieve better ROC area coverage. But to some cases, such as: "DriveCar" and "GetOutCar", the spatial saliency models obtain better performance. One sample of the video sequences from "DriveCar" category with fixations is demonstrated in Fig. 5(a). From this video, we can find that the movement of the objects out of the window is much more conspicuous than the movement inside of the car. But this kind of movements is not the focus of viewers. Therefore, in this case, the motion saliency map shown in Fig. 5(b) cannot cover the subjects' fixations.

We also combine some representative spatial saliency detection models with different temporal saliency detection models and then compare with the proposed STAM. The average ROC area comparison in every category is shown in Fig. 6. Compared with the results shown in Table 4, it can be found that almost all the integrations are better than any separated one. STAM reaches the best ROC area coverage. Furthermore, in this dataset, one example of the video saliency detection results with fixations of our proposed STAM is demonstrated in Fig. 7. In 2nd frame, the man stands quietly and looks outside from the window. The spatial saliency map is dominant and obtains good fixations coverage. Outside the window, we find a boy with white coat is running along the road. Therefore, the temporal saliency map is dominant in attention model from the 17th frame to the 55th frame, and our output attention map has peak salient value in the body of the running boy.

We also combine some representative spatial saliency detection models (with different temporal saliency detection models and then compare with the proposed STAM. The average ROC area comparison in every category is shown in Fig. 6. Compared with the results shown in Table 4, it can be found that almost all the integrations are better than any separated one. STAM reaches the best ROC area coverage. Furthermore, in this dataset, one example of the video saliency detection results with fixations of our proposed STAM is demonstrated in Fig. 7. In 2nd frame, the man stands quietly and looks outside from the window. The spatial saliency map is dominant and obtains good fixations coverage. Outside the window, we find a boy with white coat is running along the road. Therefore, the temporal saliency map is dominant in attention model from the 17th frame to the 55th frame, and our output attention map has peak salient value in the body of the running boy.

Table 4. ROC area comparison of spatial saliency models/temporal saliency models on the eye-tracking action videos

ROC Area	Spatial Saliency Models			Temporal Saliency Models		
	VOIS	CAL_VOIS	GBVS	DCOF	COF	SOF
AnswerPhone	**0.5152**	0.5030	0.4979	**0.6098**	0.5303	0.5910
DriveCar	**0.6126**	0.6050	0.6076	**0.5233**	0.4817	0.5195
Eat	**0.6310**	0.6278	0.6246	**0.6902**	0.6598	0.6644
FightPerson	**0.5057**	0.4913	0.4959	**0.6045**	0.5535	0.6005
GetOutCar	**0.5751**	0.5720	0.5707	**0.5260**	0.4874	0.5212
HandShake	**0.4730**	0.4720	0.4716	**0.6993**	0.6485	0.6934
HugPerson	**0.4625**	0.4610	0.4544	**0.6402**	0.5602	0.5996
Kiss	**0.4912**	0.4855	0.4787	**0.5833**	0.5120	0.5503
Run	**0.5469**	0.5360	0.5268	**0.5535**	0.5104	0.5496
SitDown	**0.5999**	0.5946	0.5862	**0.5183**	0.4761	0.5074
SitUp	**0.4840**	0.4854	0.4822	**0.5171**	0.4871	0.5006
StandUp	**0.6628**	0.6594	0.6530	**0.5602**	0.5269	0.5601

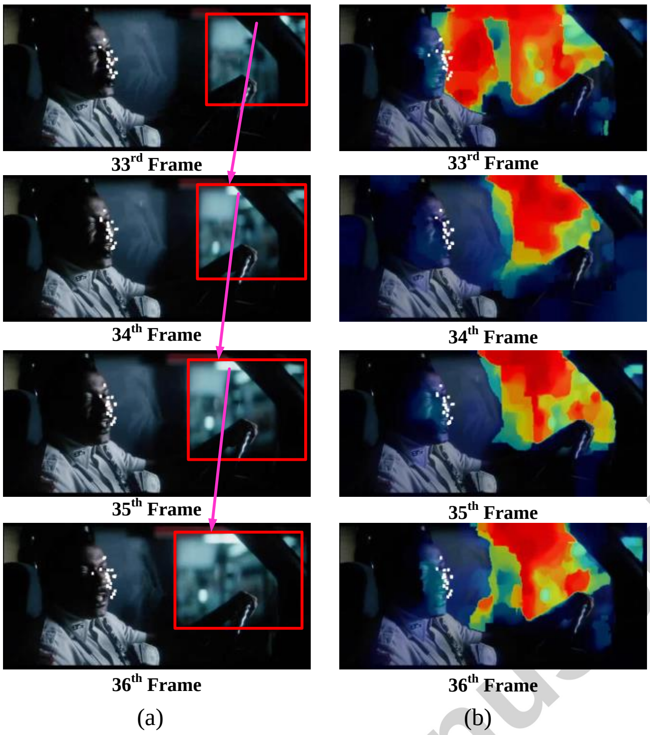


Fig. 5. (a) One sample of the video sequences from “DriveCar” category with fixations (white dots), (b) Temporal saliency map visualization to the corresponding video frame with fixations. The red box represents the movement in the current frame.

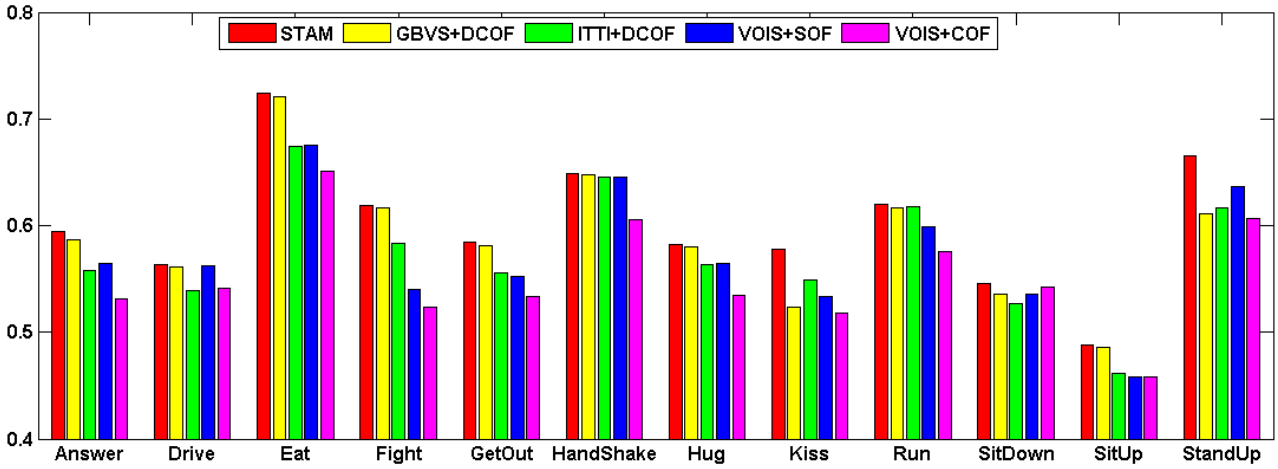
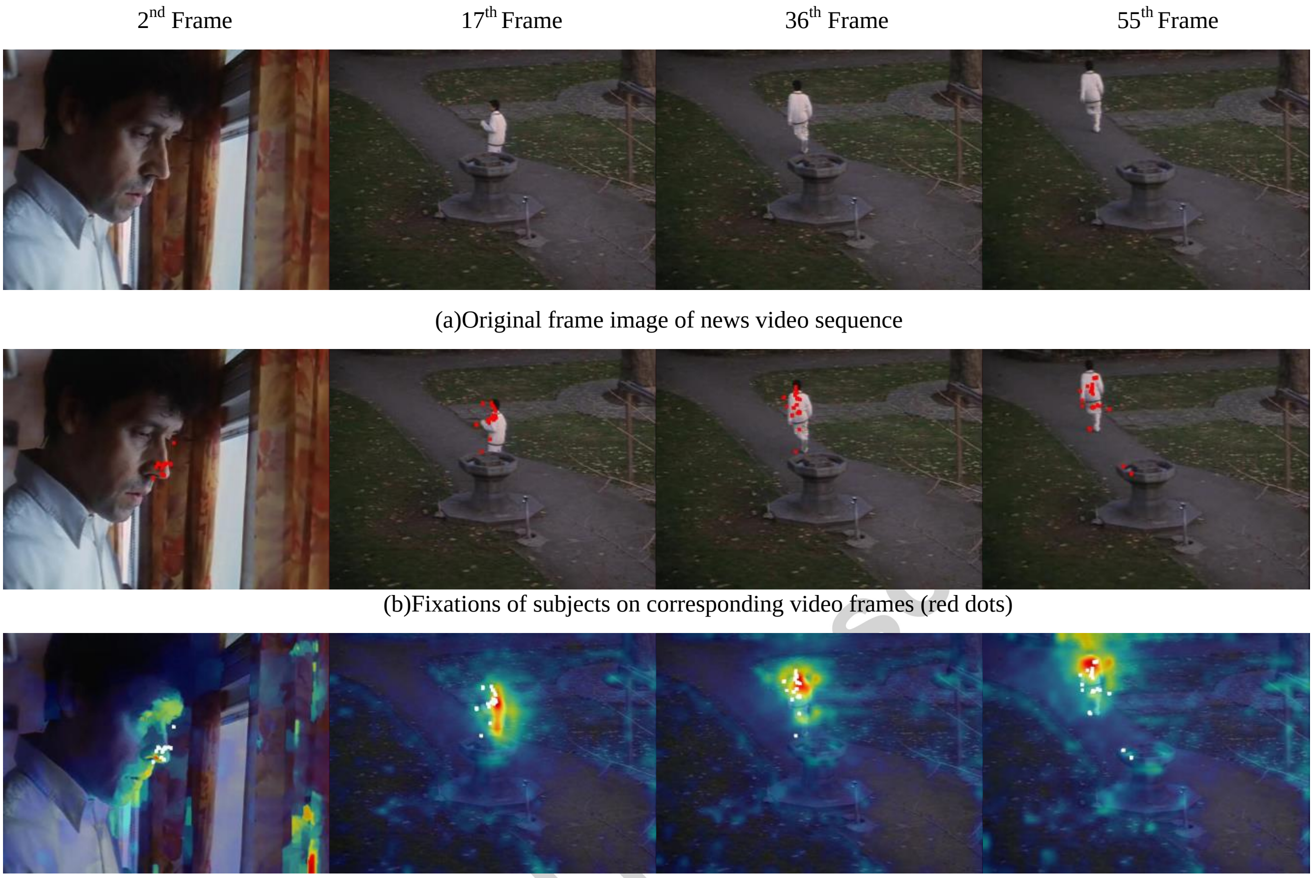


Fig. 6. Average ROC area comparisons of different combination of spatial saliency map and temporal saliency map in all categories.



(a)Original frame image of news video sequence

(b)Fixations of subjects on corresponding video frames (red dots)

(c)Saliency map overlaid of STAM with fixations (white dots)

Fig. 7. Example of video saliency detection results with fixations.

5. Conclusions

This paper proposes a novel spatial-temporal saliency detection model for video saliency detection. In spatial saliency detection, we follow three stages of the classical bottom-up spatial saliency features. We propose a human-like orientation feature map extraction based on the visual orientation inhomogeneity of human perception, and combine the orientation feature with other extracted low-level features as feature maps. In temporal saliency detection, a novel optical flow model is proposed based on the dynamic consistency of motion. Two major advantages of the proposed model can be achieved: (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 demonstrate the performance of the proposed spatial saliency model based on the visual orientation inhomogeneity of human perception goes beyond the existing spatial saliency models. As well as the progress in spatial saliency models, the proposed temporal saliency model also demonstrates better effectiveness and efficiency than the representative optical flow models. Experimental results clearly evidence that the extracted salient regions by the proposed spatial-temporal 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, such as video segmentation. The second direction is to propose novel perception-oriented video attention models by referring to more characters of human visual system.

6. Acknowledge

This work was supported by the National Natural Science Foundation of China (No. 61502311, No. 61373122), Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase), the Science and Technology Innovation Commission of Shenzhen under Grant (No. JCYJ20150324141711640), and the Shenzhen University research funding (201535).

References

[1] J. J. Gibson. *A theory of direct visual perception*, in J.R. Royce and W.W. Rozeboom, (eds): The Psychology of Knowing, Gordon & Breach, New York, 1972.

[2] T.-S. Ou, Y.-H. Huang, and H. H. Chen, “SSIM-based perceptual rate control for video coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, *20(11)*, pp. 682-691, May. 2011.

[3] Y.J. Chen, K. S. Wu, Q. Zhang, “From QoS to QoE: A tutorial on video quality assessment,” IEEE Communications Surveys & Tutorials, vol. 17, no. 2, pp. 1126-1165, 2015.

[4] H. Qu, X.R. Xie, Y.S. Liu, M.L. Zhang, L. Lu. “Improved perception-based spiking neuron learning rule for real-time user authentication,” *Neurocomputing*. 151, pp. 310-318, 2015.

[5] B.Y. Lei, S. A. Rahman, I. Song, “Content-based classification of breath sound with enhanced features,” *Neurocomputing*, 141, pp. 139-147, Oct. 2014.

[6] B.Y. Lei, I. Song, “Perception-based audio watermarking scheme in the compressed bitstream,” *AEU-International Journal of Electronics and Communications*, 69(1), pp. 188-197, Jan. 2015.

[7] J.R. Anderson, *Cognitive psychology and its implications*. 6th Edition, Worth Publishers, Oct. 2009.

[8] A. Borji, L. Itti, “State-of-the-art in visual attention modeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), pp. 185-207, Jan. 2013.

[9] J. Han, L. Sun, X. Hu, J. Han and L. Shao, “Spatial and temporal visual attention prediction in videos using eye movement data,” *Neurocomputing*, 145, pp. 140-153, Dec. 2014.

[10] Z. Wang and Q. Li, “Information content weighting for perceptual image quality assessment,” *IEEE Transactions on Image Processing*, *20(5)*, pp. 1185-1198, May. 2011.

[11] X.J. Zhang, C. Xu, X.L. Sun, G. Baciu, “Salient object detection via nonlocal diffusion tensor,” *International Journal of Pattern Recognition and Artificial Intelligence*, 29(7), pp. 1-19, Nov. 2015.

[12] X. Zhen, L. Shao, X. Li, “Action recognition by spatio-temporal oriented energies,” *Information Sciences*, 281, pp. 295-309, Oct. 2014.

[13] Y.M. Fang, W.S. Lin, B.-S. Lee, C.-T. Lau, Z.Z. Chen, and C.-W. Lin. “Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum,” *IEEE Transactions on Multimedia*. *14(1)*, pp. 187-198, Feb. 2012.

[14] J. Han, K. Li, L. Shao, X. Hu, S. He, L. Guo, J. Han, T. Liu, “Video abstraction based on fMRI-driven visual attention model,” *Information Sciences*, 281, pp. 781-796, Oct. 2014.

[15] J. Li, D. Xu, W. Gao, “Removing label ambiguity in learning-based visual saliency estimation,” IEEE Transactions on image processing, 21(4), pp. 1513-1525, Apr. 2012.

[16] L. Itti and P. Baldi, “Bayesian surprise attracts human attention,” *Vision Research*, *49(10)*, pp. 1295-1306, Jun. 2009.

[17] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *Proc. ACM MM*, pp. 815-824, 2006.

[18] C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the Underlying Neural Circuitry,” *Human Neurobiology*, *4(4)*, pp. 219-227, 1985.

[19] D. Parkhurst, K.Law, and E. Niebur, “Modeling the role of salience in the allocation of overt visual attention,” *Vision Research*, *42(1)*, pp. 107-113, Jan. 2002.

[20] S. Mathe and C. Sminchisescu, “Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition,” in *Proc. ECCV*, pp. 842-856, 2012.

[21] S.-H. Zhong, Y. Liu, F. Ren, J. Zhang, T. Ren. “Video saliency detection via dynamic consistent spatio-temporal attention modelling”. In *Proceedings of 27th AAAI International Conference on Artificial Intelligence (AAAI’13),* 2013.

[22] M. W. Jian, K.-M. Lam, J. Y. Dong, L.L. Shen, “Visual-patch-attention-aware saliency detection,” IEEE Transactions on Cybernetics, vol. 45, no. 8, pp. 1575-1586, 2015.

[23] B. Li, W.H. Xiong, W.M. Hu. “Visual saliency map from tensor analysis,” in *Proc. AAAI*, pp. 1585-1591, 2012.

[24] R.T. Born, J.M. Groh, R. Zhao, and S.J. Lukasewycz, “Segregation of object and background motion in visual Area MT: effects of microstimulation on eye movements,” *Neuron*, *26(3)*, pp. 725-734, Jun. 2000

[25] W-H. Cheng, W-T. Chu, J-H. Kuo, and J-L. Wu, “Automatic video region-of-interest determination based on user attention model,” in *Proc. ISCAS*, pp. 3219-3222, 2005.

[26] R.J. Peters and L. Itti, “Applying computational tools to predict gaze direction in interactive visual environments,” *ACM Transactions on Applied Perception*, 5(2), pp. 1-19, May. 2008.
[27] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué, “Modelling spatio-temporal saliency to predict gaze direction for short videos,” *International Journal of Computer Vision*, *82(3)*, pp. 231–243, May. 2009.
[28] L. Itti and C. Koch, “A saliency-based search mechanism for overt and covert shifts of visual attention,” *Vision Research*, 2000.
[29] M.Wang, J. Li, T. Huang, Y. Tian, L. Duan, L., and G. Jia. “Saliency detection based on 2D log-gabor wavelets and center bias,” in *Proc. ACM MM*, 2010.
[30] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20(11)*, pp. 1254-1259, Nov. 1998.
[31] R. Achanta, S. Hemami, F. Estrada and S. Süsstrunk, “Frequency-tuned salient region detection”, in *Proc. CVPR*, 2009.
[32] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Proc. NIPS*, pp. 545-552, 2007.
[33] C.C. Loy, T. Xiang, S.G. Gong, “Salient motion detection in crowded scenes,” in *Proc .ISCCSP*, pp. 1-4, 2012.
[34] B. Horn and B. Schunck, “Determining optical flow,” *Artificial Intelligence*, *17*, pp. 185–203, 1981.
[35] M. J. Black and P. Anandan, “The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields,” *Computer Vision and Image Understanding*, *63(1)*, pp. 75–104, Jan. 1996.
[36] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, “An improved algorithm for TV-L^1 optical flow,” In D. Cremers, B. Rosenhahn, A.L. Yuille, and F.R. Schmidt (Eds.), *Lecture notes in computer science, 5604, Statistical and geometrical approaches to visual motion analysis*, pp. 23-45, Berlin: Springer, 2008.
[37] D. Sun, S. Roth, M. J. Black, “Secrets of optical flow estimation and their principles,” in *Proc. CVPR*, pp. 2432-2439, 2010.
[38] D.H. Hubel, T.N. Wiesel, “Receptive fields of single neurons in the cat’s striate cortex,” *Journal of Physiology*, *148(3)*, pp. 574-591, Oct. 1959.
[39] L. Itti, C. Koch, “Computational modelling of visual attention,” *Nature Reviews Neuroscience.* 2(3), pp. 194-203, Mar. 2001.
[40] J. Thiem, C. Wolff, and G. Hartmann, “Biology-inspired early visual system for a spike processing neurocomputer,” *Biologically Motivated Computer Vision*, *1811*, pp. 387-396, 2000.
[41] B. Alexe, T. Deselaers, V. Ferrair. “What is an object? ” in *Proc. CVPR*, pp. 73-80, 2010.
[42] A.R. Girshick, M.S. Landy, and E.P. Simoncelli, “Cardinal rules: visual orientation perception reflects knowledge of environment statistics,” *Nature Neuroscience*, *14*, pp. 926-932, Apr. 2011.
[43] B. Li, M.R. Peterson, and R.D. Freeman, “Oblique effect: a neural basis in the visual cortex,” *Journal of Neurophysiology*, pp. 204-217, Feb. 2003.
[44] M. Brecht, and J. de Saiki, “A neural network implement of a saliency map model,” *Neural Networks*, 19, pp. 1467-1474, Dec. 2006.
[45] A. Oliva, and A. Torralba, “Modeling the shape of the scene: a holistic representation of the spatial envelope,” *International Journal of Computer Vision*, *42(3)*, pp. 145-175, Jan. 2001.
[46] A. Torralba, and A. Oliva, “Statistics of nature categories,” *Network: Computation in Neural Systems*, *14(3)*, pp. 391-412, May. 2003.
[47] S. H. Zhong, Y. Liu, Q.C. Chen, “Visual orientation inhomogeneity based scale-invariant feature transform,” *Expert Systems with Applications*, 42(13), Nov. 2015.
[48] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8(6)*, pp. 679-698, Nov. 1986.
[49] D.H. Warren and E.R. Strelow. *Electronic spatial sensing for the blind: contributions from perception*. Martinus Nijhoff Publishers, Massachusetts, 1985.
[50] Gibson, J.J. *The Perception of the Visual World*. Houghton Mifflin, 1950.
[51] C. Liu, P.C. Yuen, G.P. Qiu, “Object motion detection using information theoretic spatio-temporal saliency,” *Pattern Recognition*, *42(11)*, pp. 2897–2906, Feb. 2009.
[52] M. Carrasco and K. Frieder, “Cortical magnification neutralizes the eccentricity effect in visual search,” *Vision Research*, *37(1)*, pp. 63–82, Jan. 1997.
[53] E. Vul, M.C. Frank, J.B. Tenenbaum, G. Alvarez. “Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model,” in *Proc. NIPS*, pp. 1955-1963, 2009.
[54] R. Milanese, H. Wechsler, S. Gil, J.-M. Bost and T. Pun. “Integration of bottom-up and top-down cues for visual attention using non-linear relaxation,” in *Proc. CVPR*, pp. 781-785, 1994.
[55] A. Rahman, D. Houzet, D. Pellerin, S. Marat, and N. Guyader, “Parallel implementation of a spatio-temporal visual saliency model,” *Journal of Real-Time Processing*, 6, Special Issue, pp. 3-4, 2011.
[56] H. J. Seo, P. Milanfar, “Static and space-time visual saliency detection by self-resemblance,” *The journal of vision*, *9(12)*, pp. 1-27, Nov. 2009.
[57] X.D. Hou, J. Harel, and C. Koch, “Image signature: highlighting sparse salient regions,” IEEE Trans. *Pattern Analysis and Machine Learning*, 34(1), pp. 194-201, Jan. 2012.
[58] T. Judd, K. Ehinger, F. Durand, and A. Torralba. “Learning to predict where humans look,” in *Proc. ICCV*, 2009. pp. 2106-2113.
[59] J. Han, S. He, X. Qian, D. Wang, L. Guo, T. Liu, “An object-oriented visual saliency detection framework based on sparse coding representation,” IEEE Trans. on Circuits and Systems for Video Technology, 23(12), pp. 2009-2021, Dec. 2013.
[60] P. Viola and M. Jones, “Robust real-time object detection,” in *Proc. SCTV*, 2001.
[61] J. Li, Y. Tian, T. Huang, and W. Gao, “Cost-sensitive rank learning from positive and unlabeled data for visual saliency estimation,” *IEEE Signal Process Letter*, *17(6)*, pp. 591-594, Jun. 2010.

Sheng-hua Zhong received her B.Sc. in Optical Information Science and Technology from Nanjing University of Posts and Telecommunication in 2005 and M.S. in Signal and Information Processing from Shenzhen University in 2007. She received her Ph.D. from Department of Computing, The Hong Kong Polytechnic University. From 2013 to 2014, she worked in Department of Psychological & Brain Sciences from Johns Hopkins University as a Postdoctoral Research Associate. She is currently a lecturer in College of Computer Science & Software Engineering from Shen Zhen University. Her research interests include vision science, multimedia content analysis and cognitive modeling.

Yan Liu received the BEng degree from the Department of Electronic Engineering at Southeast University and the MSc degree from the School of Business at Nanjing University in China. She received the PhD degree from the Department of Computer Science at Columbia University. Currently, she is an associate professor in the Department of Computing at The Hong Kong Polytechnic University. As a director of Cognitive Computing lab, she focuses her research in brain modeling, ranging from image/video content analysis, music therapy, manifold learning, deep learning, and EEG data analysis.

Vincent Ng is an Associate Professor in the Department of Computing, The Hong Kong Polytechnic University. His research interests include social media analysis, learning analytics, and medical image analysis. Dr. Ng is also active in consultancy service for companies, government agencies and voluntary associations. He was a board member of the Public Examination Board of the HKEAA and worked with many local schools in Hong Kong.

Yang Liu received the BS and MS degrees in automation from the National University of Defense Technology in 2004 and 2007, respectively. He received the PhD degree in computing from The Hong Kong Polytechnic University in 2011. Between 2011 and 2012, he was a postdoctoral research associate in the Department of Statistics at Yale University. He is currently a research assistant professor in the Department of Computer Science at The Hong Kong Baptist University. His research interests include cognitive science, machine learning, applied mathematics, as well as their applications in brain modeling, high-dimensional data mining, visual content analysis, and music therapy.

*Photo of the author(s)

*Photo of the author(s)

*Photo of the author(s)

*Photo of the author(s)