Looking Into Saliency Model via Space-Time Visualization

Haoran Liang, Ronghua Liang, *Member, IEEE*, and Guodao Sun

***Abstract*—We introduce a visual analytics method to analyze eye-tracking data and saliency models for dynamic stimuli, such as video or animated graphics. The focus lies on the analysis of the different performance of saliency models in contrast to human observers to identify trends in the general viewing behavior, including time sequences of attentional synchrony and objects with a strong attentional focus. By using a space-time cube visualization in combination with clustering, the dynamic stimuli and associated eye gazes as well as the attention maps from saliency models can be analyzed in a static three-dimensional representation. We propose algorithms to keep the appearance of the computer's attention data in line with the human's eye-tracking data. The analytical process is supported by multiple coordinated views that allow the user to focus on different aspects of spatial and temporal information in eye gaze data and saliency map. By comparing attention data from both human and computer incorporated with the spatiotemporal characteristics, we are able to find the different patterns within human and computer algorithms. We list our key findings to help developing better saliency detection algorithms.**

***Index Terms*—Saliency model, spatiotemporal analysis, visualization.**

I. Introduction

HUMANS' tremendous ability to rapidly direct gaze and select the most relevant information from the visual world around us have been intensely researched for years. Understanding and simulating this attentional mechanism has both scientific and economic impact, and are attracting increasing attention both human vision and computational vision [1]–[3]. Besides the traditional mathematical metrics to explore visual attention, the development of visualization techniques nowadays allows us to deeply look into the gap between human and computer, which suggest a better design and improvement of artificial intelligence in the human cognitive domain.

The last decade has witnessed the rapid development of saliency detection techniques. A number of saliency detection models have been proposed to simulate human's behavior and explore the most attractive parts in static or dynamic stimulus. Several metrics are being used to measure the performances of different models quantitatively. However, the rank of metric scores of different saliency models does not intuitively tell where the differences exist and what lead to them. Due to the processing method and noise, generally people can barely judge and select saliency map that has the highest metric score among the results obtained by different saliency models. Therefore a new way to explore and analyze the saliency models and eye-tracking data is needed for a better and deeper understanding of visual saliency domain and human cognitive processes.

The main focus in visual saliency research in the past lies on the analysis of static stimuli such as images. For visualizing fixations, numerous methods such as heat map and scan path are widely used on data recorded for static stimuli. For the analysis of dynamic stimuli such as video sequences, however, the number of available visualization methods is very limited and often, those techniques are not very effective because of: 1) In general, the analysis of eye-tracking data from dynamic stimuli can be achieved by watching the video with afore mentioned methods, which is a time-consuming and exhausting task for the analyst. 2) Statistical analysis of AOIs requires either a reliable detection algorithm for locating, or tedious manual editing.

Future improvements in the field of computer vision may provide techniques that can successfully identify the objects as well as the regions of interest. However, human observation is still required for semantic interpretation. Moreover, for analyst, it would be more efficient to look at a representation of the whole video at once and find the important clips that contain interesting features without a sequential search through each frame.

The existing method of visualizing eye-tracking data only consider human's behavior while in the field of building saliency detection model, a more important thing is to explore the different observation patterns between human and computer when being presented visual stimulus. In this paper, we propose visualization designs for the purpose of analyzing saliency models and eye-tracking data intuitively. The main contributions are listed as follows:

1) We for the first time visualize both eye-tracking data from human and saliency data from saliency detection model in a single system for a better understanding and exploration of the different observation patterns between human and computer algorithm. (See Fig. 1)
2) Our approach can provide a spatial and temporal view of visual attention data from both human and computer, which allows a wide collection of different analysis tasks.
3) We list our key findings from our analysis on human and computer's attention data, suggesting better design of saliency detection algorithm and showing deeper understanding of human observation pattern.

Manuscript received May 9, 2016; revised August 8, 2016; accepted September 21, 2016. Date of publication September 26, 2016; date of current version October 19, 2016. This work was supported in part by the National Science Foundation of China under Grant 61527808, Grant 61379076, and Grant 61602409; and in part by the Outstanding Youth of Zhejiang Provincial Natural Science Foundation under Grant LR14F020002. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Nan Cao.

The authors are with the College of Information Engineering, Zhejiang University of Technology, Hangzhou 310013, China (e-mail: yanakz@zjut.edu.cn; rhliang@zjut.edu.cn; guodao@zjut.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2016.2613681

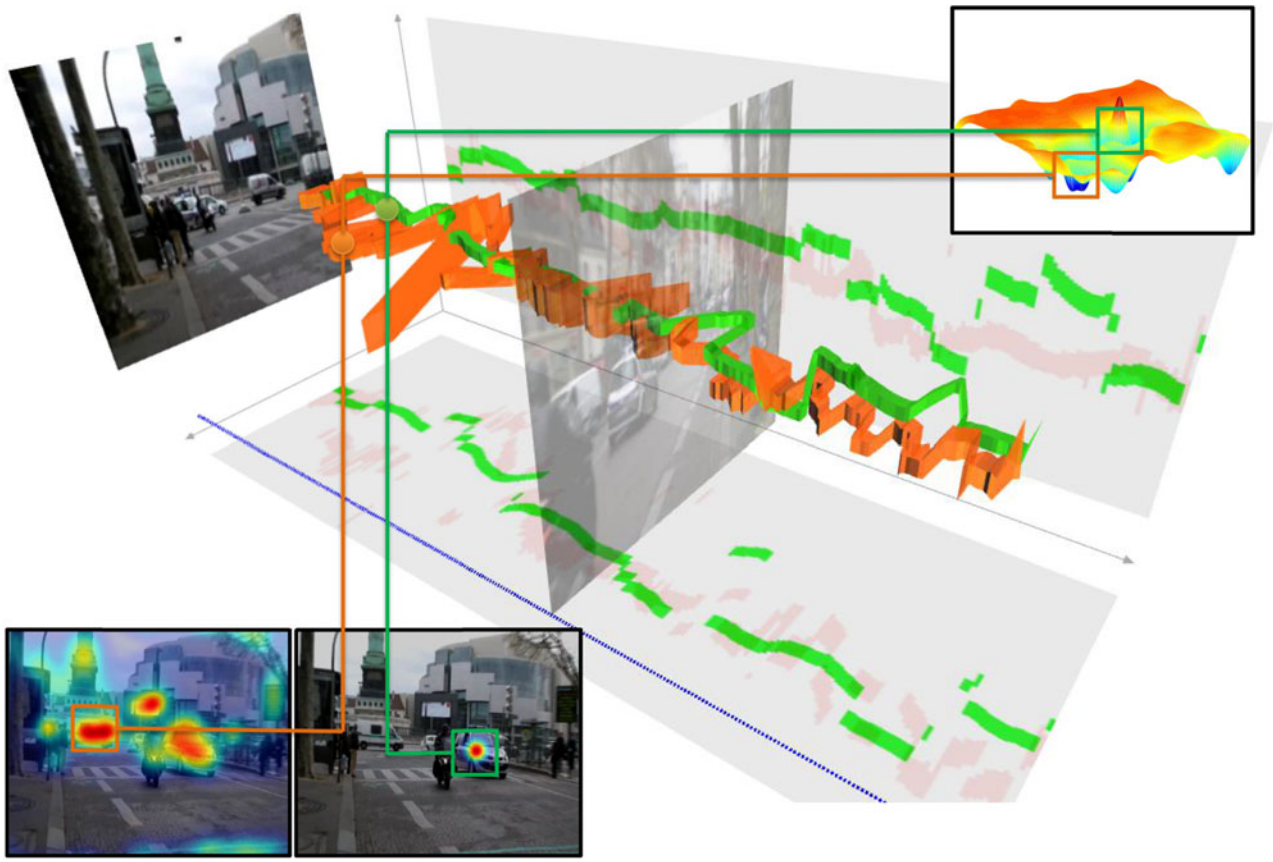


Fig. 1. Visualization of area of interest from both human (green cubes) and saliency model (red cubes) for a video stimuli.

II. Related Work

A. Visual Saliency

Recent years have witnessed the increasing interest in the fields of both psychology and computer vision [2], [4]–[7]. Computational saliency models predict important locations of a visual scene and focus limited resources to the identified regions [1], [8]–[11]. The first saliency model was proposed by Koch and Ullman [2] and later implemented by Itti *et al.* [12], inspired by which, a number of algorithms have been developed to predict where humans look at in images along the same line [4], [13]. In these models, low-level features such as color, intensity and orientation were extracted and feature channels were computed through center-surround filtering at multiple spatial scales, followed by a feature integration step using linear mechanism to obtain the saliency map.On the basis of the effectiveness of color, intensity and orientation proved by many works in guiding visual search and attention-based computational model, various computational algorithms were proposed to infer saliency of different feature channels. The information maximization model [14] was based on the closely related quantity of self-information using Independent Component Analysis (ICA) decomposition [15]. Vasconcelos *et al.* demonstrated discriminant saliency with the center-surround hypotheses [16]. Spectrum-based algorithms [17], [18] were also developed to remove the redundant content of a visual scene to predict saliency regions.

With the improved integration algorithms, these models perform better than the classic approach. A more recent problem in the saliency community is the semantic gap between the predictive power of computational saliency models and human behavior. That is, pixel-level image attributes fail to encode object and/or semantic information, which is many times more important to saliency than pixel-level information. As recent psychophysical [19], [20] and computational studies [21], [22] suggest, visual attention is attracted, at least largely, by semantically interesting regions or objects, especially in complex visual scenes like web page and crowds [23], [24]. To fill the semantic gap between computational saliency models and human performance, specifically-trained object detectors have been incorporated into saliency models. For example, faces have been shown to attract attention independent of tasks, and several recent models [13], [21], [25], [26]. They combined face detection as a separate visual cue with traditional low-level features to improve saliency detection. Furthermore, Judd et al. [6] proposed a Support Vector Machine (SVM) based learning approach to linearly combine face, pedestrian and car detectors with low- and mid-level features. The integration of multiple object detectors in general boosts prediction performance, especially scenes with the objects that have the detectors built and integrated.

Whereas only spatial features are available for the saliency detection of still images, temporal features can be also exploited for the spatiotemporal saliency detection of video sequences. Some researchers have extended existing spatial saliency

detection schemes, by considering the additional temporal dimension, to extract spatiotemporal saliency [27], [28].

B. Eye Tracking Data Visualization

The use of eye-tracking can provide valuable information to understand the viewing behavior of human. Thus visualization is a vital part in visual saliency research to not only show the eye movement distribution, but also provide the probability to explore the potential pattern within human's viewing behavior. Several works have measured the gaze overlaps of a video that showed a surgical task to compare experts gaze with the gaze of trainees [29]. Some examined similarities in the viewing behavior of several users to identify centers of interest in movie scenes [30]. Marchant *et al.* [31] described an approach to investigate the influence of directorial techniques on film viewers experience. Smith and Henderson [32] compared the degree of attentional synchrony between static and dynamic scenes.

There exist a number of methods to visualize eye-tracking data. Holmqvist *et al.* [33] provide a comprehensive guide to methods and measures. Generally, heat maps [34]–[36] are used to display aggregated eyetracking data. Tsang *et al.* [37] provide a tree-like visualization for the exploration and comparison of sequential gaze orderings. Raschke *et al.* [38] introduced the parallel scan-path visualization to facilitate the comparison of eye-tracking data between several users. In the context of visual analytics, Andrienko *et al.* [39] provide a detailed methodology for eye movement data. We adopt many of the standard visualization methods in our work.

Space time cube is widely used in various fields of research. Gatalsky *et al.* [40] describe its application to event data in a geographical context. Chen *et al.* [41] and Botchen *et al.* [42] represent video content in 3D to depict individual objects and motion events. In the context of eye-tracking, Li *et al.* [43] describe the use of the space-time cube to visualize eye-trajectories. They focus on the analysis of static stimuli. For the application to dynamic stimuli, Duchowski and McCormick [44] describe a space-time representation of Volumes Of Interest for aggregated eye movement trajectories. Kuno and Daniel [45] extend the concept for dynamic stimuli and provide different data representations in addition to the mentioned eye-trajectories. Clustering of eye-tracking data is already used when fixations are identified in raw data. Salvucci and Goldberg [46] describe a taxonomy for different fixation identification algorithms. For the clustering of multiple user gaze data, Sawahata *et al.* [47] and Mital *et al.* [48] use a Gaussian Mixture Model. Here in this work we use the mean-shift clustering approach for gaze data, according to Santella and DeCarlo [49] because it is robust to noise and does not require a preset number of clusters.

However, past works only focus on human's eye-tracking data, none of them have combined attention data from both human and computer algorithms (from saliency models) for analysis, which is essential to explore the gap between human's observation pattern and saliency algorithm. Therefore, it is our goal to bridge the two modalities of data and find the key difference and correlation.

III. Design Overview

In this section, we introduce the design of our visual analysis system including the multiple views of spatiotemporal eye-tracking and predicted visual saliency data, the comparison part that uses heat map to explore the differences in results from saliency model and human. We also briefly introduce the visual stimuli used in our analysis.

A. Interface

Fig. 2 shows a screenshot of our system. There are four main components:

1) *Visualization view:* The visualization view consists of two components. The main part is the interactively explorable space-time cube that visualizes the areas of interest (AOI) found in the visual attention data from human or computer algorithm. The presented data are freely rotatable and movable for investigation. Users can navigate through the video by using the time-scroll which also indicates the current frame being played. Each slice of the AOIs presented in the center of screen is selectable for the users to quickly find a particular frame.
2) *Data view:* This component allow users to import data from new saliency models. We take the predicted saliency maps as the input and calculate all the needed data such as the heat map, AOI and metric score. The data that are enabled in the list will be presented in the visualization view, so the users can navigate multiple models' data in one coordinate system to explore the difference and correlation. Additionally, we allow users to modify color and transparency to customize the 3-D view of eye-tracking data.
3) *Saliency map and heat map view:* In this component we show saliency maps and difference heat maps from the given data. The definition and detail of 2-D and 3-D difference heat map are described in Section V-C.
4) *Statistics view:* This component lists all the metrics named AUC, CC and NSS for the given stimuli. We also include the quantitative measure scores for main factors. The definitions of the investigated factors are described in Section VI-A.

B. Stimuli

We use two databases in our work, the first one is a public database named Coutrot Database [50], another one is synthesized using a collection of static images.

1) Coutrot Database: The visual material consisted of 15 one-shot conversation scenes extracted from French Hollywood like movies. Videos featured two to four conversation partners embedded in a natural environment. Videos lasted from 12 to 30s, had a resolution of 720×576, and a frame rate of 25 frames per second. The stimuli features conversation partners embedded in complex scenes (cafe, streets, corridor, office, etc.) involving different moving objects (glasses, spoons, cigarettes, papers, etc.). The database also contains auditory material consisted of 45 monophonic soundtracks: a first set of 15 soundtracks extracted from the conversation scenes (dialogues), a

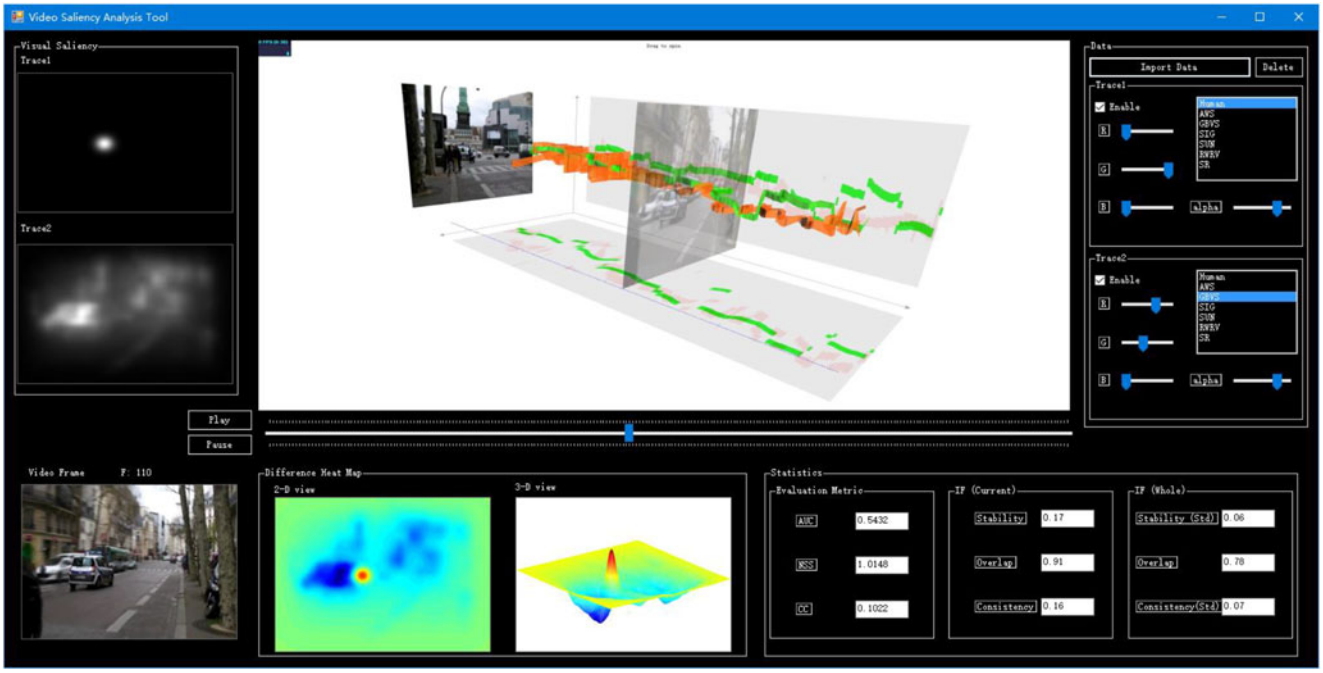


Fig. 2. Overview of the design components.

second set of 15 soundtracks made up of noises from moving objects (short abrupt onsets, e.g., falling cutlery), and a third set of 15 soundtracks extracted from landscape scenes (continuous auditory stream, e.g., wind blowing).

There are four versions of eye-tracking data that are recorded using the same video with different soundtracks. In order to focus more on the visual features, we only use the eye-tracking data recorded using the original soundtrack in this work.

2) Synthesized Stimuli: We collected a total of 300 natural images, representing a variety of common scene from Flickr. Each time, we randomly select 5 images to synthesize a video that lasts for 15 s, each image is presented for 2.5 s followed by a gradient transition (0.5 s) to the next one, forming 10 clips of synthesized videos. Fifteen students (8 male and 7 female, between the ages of 18 and 25) with corrected or uncorrected normal eyesight free-viewed the full set of videos. These videos were presented on a 22-inch LCD monitor (placed 57 cm from the subjects), and eye movements of the subjects were recorded using an Eyelink 1000 (SR Research, Os-goode, Canada) eye tracker, at a sample rate of 1000Hz. The screen resolution was set to 1680×1050, and the images were scaled to occupy the full screen height when presented on the display. Therefore, the visual angle of the stimuli was about $38.8^\circ \times 29.1^\circ$, and each degree of visual angle contained about 26 pixels in the image.

IV. Saliency Visualization

In our visual analysis approach, we included established visualization methods for eye-tracking data, namely Heat Map and Areas of Interest (AOI). The two methods are commonly known and used by researchers, which allows for an easy adoption of design. Moreover, the methods can express visual attention data intuitively, making it more understandable in a wide collection of different analysis tasks.

Fig. 3. Areas of interest obtained by human (left) and saliency model (right). Generally, AOI predicted by accurate saliency model will contain human's AOI to a certain extent.

A. Heat Map

Heat map is the most common method in visual saliency research to provide a qualitative view of observers' gaze distribution. It convolves fixation map with a Gaussian, i.e., for each fixation, it adds a 2D-Gaussian centered at that location. Especially for static stimuli, given the full set of fixations, heat map could completely express the aggregation of gaze positions over the observation time. The principle of static heat maps can be applied to dynamic stimuli to summarize the dtribution of attention, but the duration for observing each frame is quite short, leaving no more than one fixation for that frame (See Fig. 3). Most current methods integrate static saliency approaches with spatial and temporal factors to predict video saliency, resulting in highlighting multiple regions in saliency maps, which makes it ambiguous for deciding the exact region being fixated by observers. This requires a preprocess of saliency map before comparing with human fixation.

B. Areas of Interest

Fixations usually cluster on salient visual references (Fig. 7), which denote regions that might be of interest for viewers and analyst. Common eye movement metrics such as fixations per AOI or percentage of participants fixating an AOI can be used to retrieve objective information. Given eye-tracking data of video material, unknown AOI can be found by applying clustering algorithms. In our visual analysis approach, we find AOIs based on saliency maps. The principle to extract AOI directly from saliency map can be achieved by selecting salient regions that are highlighted after utilizing saliency prediction models. The selection detail is described in Section V-D.

C. Space Time Cube Visualization of Dynamic Saliency Changes

The areas of Interest view and heat map view only present a static display of the predicted saliency map as well as the difference between the predicted saliency map and ground truth. However, if a user needs to examine the saliency map of different frames or the dynamic difference between the predicted ones and ground truth, he or she may need to examine different views one by one, and try to correlate them together. This may lead to cognitive burden, and low efficiency of task completion. In order to overcome the issues above, we employ space-time cube visualization to present and analyze the overall dynamic changes of the saliency map as well as the comparison between them. Space-time cube is commonly used to analyze temporal dynamic, space related data including eye-tracking data. In Fig. 1, cubes at each time step represent the bounding box of salient area with red encoding the predict ones and green encoding the ground truth by human. The cubes are stacked along the x-axis to reveal the overall dynamic changes and difference of the saliency map. The cubes representing predicted ones and ground truth at a same time are overlapped with alpha blending to ensure both of them could be perceived visually.

The space-time cube visualization supports basic interactions including zooming and rotating, and advanced interactions including linking with heat maps view and area of interest. Users could click any single frame to examine the detailed difference of the predict saliency map and the ground truth.

V. Data Analysis

In this section, we first briefly list the saliency models we use in our visualization system. Secondly, we introduce the process to generate the 2-D and 3-D difference heat map that intuitively show where the differences exist. Lastly, we propose our method to locate AOI when given a static saliency map.

A. Salieny Models

We select six state-of-the-art saliency models that are purely bottom-up in our work, namely the Adaptive Whitening Saliency (AWS) [51], the Graph Based Visual Saliency (GBVS) [52], the Image Signature (SIG) [17], the Saliency Using Natural Statistics (SUN) [53], Spatiotemporal Saliency Detection for Video Sequences Based on Random Walk With Restart (RWRV) [27], Visual Saliency Detection by Self-Resemblance, (SR) [28]. Among them, the first four models were primarily developed for static stimuli while the last two were for video data by taking spatial and temporal factors into consideration. To apply static saliency model for video data, we simply input each frame independently to generate saliency map.

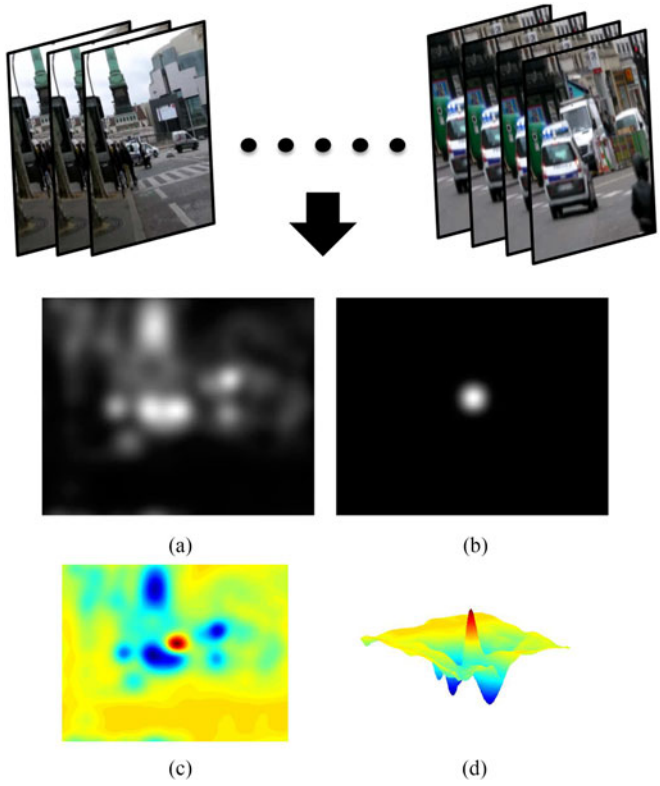


Fig. 4. Illustration of the difference between predicted saliency map and ground truth. (a) Saliency map obtained by adaptive whitening saliency (AWS). (b) Ground truth by human. (c) The 2D view of the difference heat map. (d) The 3D view of the difference heat map.

B. Map Preprocessing

The saliency maps obtained by different models are firstly blurred using Gaussian with a same standard deviation σ that varies from 0 to 10. After that we convert saliency maps to binary images with a threshold δ, i.e., we only keep the pixels that have salient values larger than δ and set the rest to 0. Let p be the pixel value in the saliency map, the above approach can be written as

$$B(p) = \begin{cases} 1, & G(p) \geqslant \delta \\ 0, & G(p) < \delta \end{cases} \tag{1}$$

where $G(\cdot)$ denotes the Gaussian function. In our work, we set δ to 0.8 as it satisfyingly filters the non-salient parts of image and only keeps no more than 3 regions for each frame.

C. Difference Heat Map

Traditional saliency prediction methods use metrics such as shuffled Area Under Curve (sAUC), Normalized Scanpath

Saliency (NSS) and Linear Correlation Coefficient (CC) to quantitatively evaluation the performance. Additionally, heat maps are usually overlaid on top of the original stimulus or displayed standalone for viewers to compare by themselves. Here in our visual analysis approach, we propose using the difference heat map to intuitively show the way human and saliency model observe as well as the pattern in finding salient regions.

A slice inside in Fig. 4 represents the current video frame. Its corresponding saliency maps by human [Fig. 4(b)] and saliency model [Fig. 4(a)] are shown alongside. Its position is freely rotatable and movable for the users to investigate data around it. Fig. 4(c) is obtained from the subtraction of Fig. 4(a) from 4(b). To ensure the two maps have the same data range, they are normalized by the maximum value inside the map. Let p be the set of pixel value in saliency map, $g(p)$ and $m(p)$ be the saliency maps for human and saliency model respectively. The process of generating difference heat map $h(p)$ can be written as

$$h(p) = \frac{g(p)}{\max(g(p))} - \frac{m(p)}{\max(m(p))}.$$

We then use each pixel value in $h(p)$ as the depth to plot colored parametric mesh in 3-D space [Fig. 4(d)]. We can conclude that the region in red and blue indicate the false positives and false negatives respectively.

D. Visual Attention Cluster Analysis

Since the eye-tracking data from human and saliency prediction results from saliency models are in different modalities, it is required to locate the AOIs using different methods.

1) Eye-Tracking Data: For eye-tracking data, the common practice to find areas that attract attention is using clustering algorithms. Those algorithms should satisfy one basic requirement: the unknown number of clusters. Due to the unpredictability of salient features such as varying color, moving object, the number of clusters can not be pre-defined properly. Even if these factors are known, the number of participants and length of stimulus will also become influence factors. In addition, the algorithms are better to be parameterizable in order to define the granularity of the clusters. In this work, we utilize Mean Shift that is widely used in the field of machine learning and computer vision to cluster eye-tracking data for it ideally fits the requirement.

The results on coutrot database show high consistency of human's behavior towards the same stimuli, resulting in one AOI for each stimuli.

2) Predicted Saliency Map: Oppositely, saliency maps obtained by saliency models usually contain three or more highlighted regions, part of which is redundant and makes it ambiguous for deciding the main attractive region in a particular frame. As described in Section V-B, we have pre-processed saliency maps using (1), the remaining of image content are several connected components, which are shown in Fig. 5. We then select n components sorted by area as the representative AOIs. In our case, we set n to 3.

We observe that in most case, the largest component in one frame fits well with human's attention data, yet we still need to fine-tune the selection strategy because saliency models to some extent find important regions that contain salient features without telling the exact one being fixated in each frame. For each frame, we calculate each selected connected component's distance to human's AOI that discovered using Mean Shift, the closest connected component among the three will be selected.

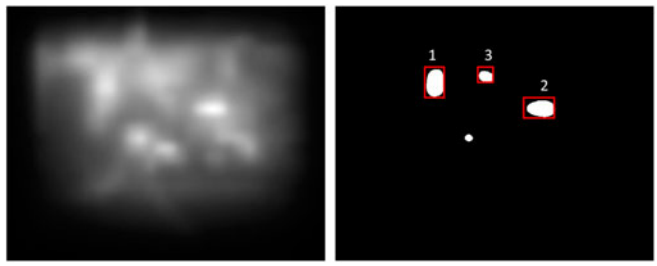


Fig. 5. Process of finding AOIs from predicted saliency map. The connected components are labeled in terms of area. Note that the maximum number of AOI n is set to three in this work.

TABLE I
QUANTITATIVE EVALUATION FOR THE SELECTED SALIENCY MODEL

Database	Coutrot [50]			Synthesized Video		
Metric	AUC	CC	NSS	AUC	CC	NSS
GBVS [52]	0.71	0.22	1.93	0.66	0.15	1.43
RWRV [27]	0.58	0.15	1.36	0.63	0.12	1.40
SR [28]	0.51	0.13	0.89	0.51	0.14	0.88
SIG [17]	0.51	0.02	0.22	0.54	0.10	0.51
SUN [53]	0.52	0.09	0.94	0.54	0.12	1.01
AWS [51]	0.67	0.20	1.74	0.63	0.18	1.63

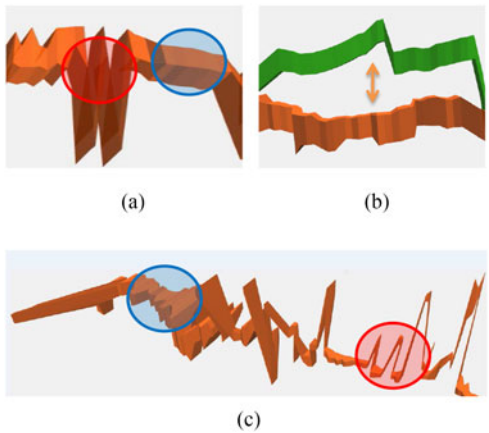


Fig. 6. Three factors that are focused on during the analysis.

TABLE II
QUANTITATIVE MEASURE FOR THE SELECTED SALIENCY MODEL

Factor	Stability	Overlap	Consistency
GBVS [52]	0.14	0.33	0.04
RWRV [27]	0.43	0.18	0.18
SR [28]	0.44	0.20	0.09
SIG [17]	0.44	0.18	0.24
SUN [53]	0.29	0.24	0.05
AWS [51]	0.17	0.33	0.05

The three factors are stability, overlap and consistency.

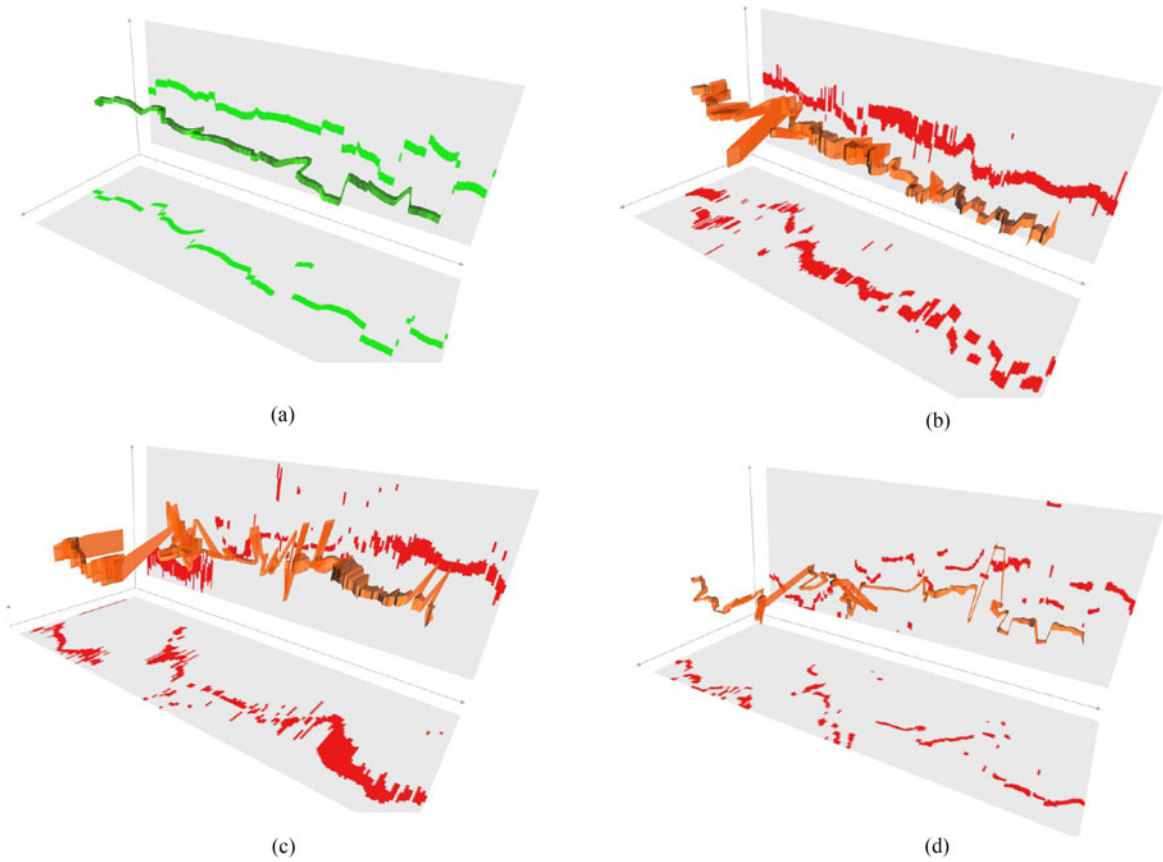


Fig. 7. 3D space-time visualization of eye-tracking data. (a) Eye-tracking data. (b) Predicted eye movement using AWS. (c) Predicted eye movement using RWRV. (d) Predicted eye movement using SR

VI. Case Study

We first list the quantitative results for each saliency models in Table. I. From the performance we can categorize the saliency models into three levels: high (GBVS [52], AWS [51]), medium (RWRV [27]) and low (SIG [17],SUN [53], SR [28]). Space-time cube visualization could help users evaluate the performance of different saliency prediction algorithms. Next we look deeply into the visualized cubes to explore the factor that leads to the difference.

A. Investigation

Three main factors are taken into consideration in our analysis. To better understand the data distribution, we also define quantitative measures for the three factors.

1) *Stability:* denotes the frequency of switching AOI. As shown in Fig. 6(a), the fixations change rapidly during the time segments in red circle yet remain still in blue circle. This could remind us of the change of salient features in the particular frames. The standard deviation of distance of adjacent AOIs in this case can be used to measure stability quantitatively.
2) *Overlap:* is used to observe the similarity of gaze distribution from human and computer algorithms. It is apparently that better results from saliency model will to the most extent overlap with eye-tracking data. The time segments in which the data do not overlap (See example in Fig. 6(b)) are valuable for investigation. Overlap can be measured by the size of overlapped region.
3) *Consistency:* refers to the size of AOI over the observation time. The AOIs obtained from eye-tracking data are not only consistent in location, but also in size. Therefore we could look into the particular frames [e.g., the time segments in red and blue circles in Fig. 6(c)] to find the reasons that lead to inconsistency. Here the standard deviation of size of AOI is used to measure Consistency of given computer model.

Particularly, the definitions of quantitative measure for Stability (S), Overlap (O), Consistency (C) of a given saliency model ($model$) can be written as

$$S(\text{model}) = std(\{d_1, d_2, ..., d_{n-1}\})$$

$$O(\text{model}) = \frac{1}{n}\sum_{1}^{n} a_i$$

$$C(\text{model}) = std(\{s_1, s_2, ..., s_n\})$$

where $std(\cdot)$ refers to standard deviation, n is the number of frame, d_i is the euclidean distance between the ith and $i+1$th AOIs from one model (the anchor is set to the centre of AOI), a_i represents the size of overlapped region of AOIs from human

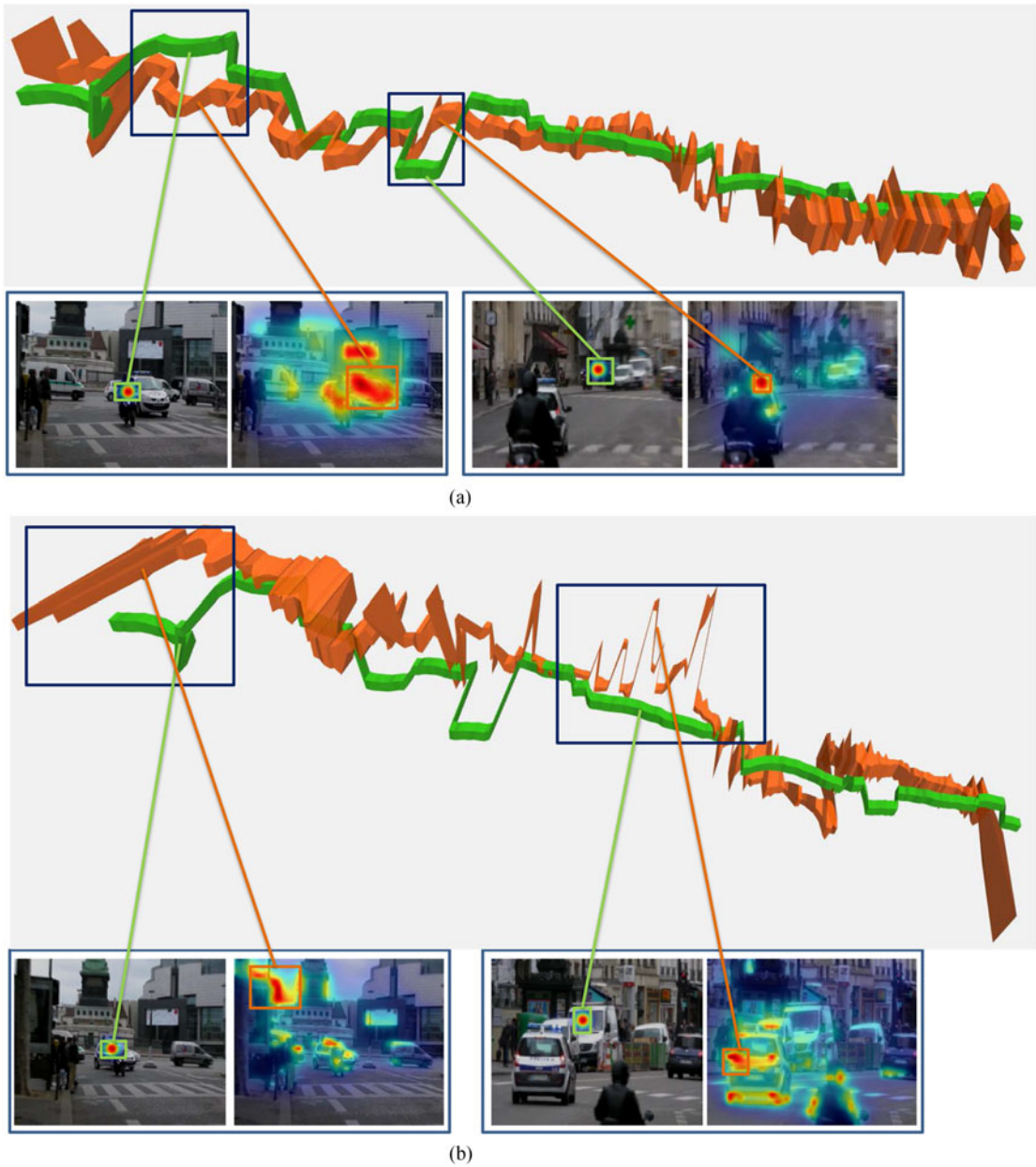


Fig. 8. Visualization of the main false detection of saliency model. (a) GBVS (red) versus human (green). (b) RWRV (red) versus human (green).

and computer model in ith frame, s_i is the size of AOI in ith frame.

We test the three measures on all the stimulus in both databases and report the average performances in Table II. Note that we have normalized the distance and size of AOI before calculate the measures for three factors, i.e., distance is normalized by the diagonal of video frame and size by the size of video frame. From the result we can see that the rank of quantitative measures approximately matches that of metric scores in Table I. However, the performance of a saliency model could not be directly judged by these three scores. For instance, let's assume d_i will always be the same but very large, the stability and size scores will approximate zero, but apparently, such model is not stable at all. Therefore, the visualization view should also be taken into consideration when measuring models.

B. Observation

The key observations are summarized below:

Stability is important in saliency prediction: Fig. 7 shows three space-time cube views for human, AWS, RWRV and SR. From top to bottom, they are sorted in descending order by the metric scores in Table. I. We can see that human's eye movement is more smooth and continuous than those obtained by saliency models. The participants usually gaze on one visual reference for more than one second before turning to another one. In contrast, eye movements predicted by saliency models rapidly jump among several regions.

Besides the frequency of switching where to look, the object that is fixated generally locates closely to the former one. This is because that people tend to focus on the regions that contain or possibly contain salient objects such as the street and vanishing

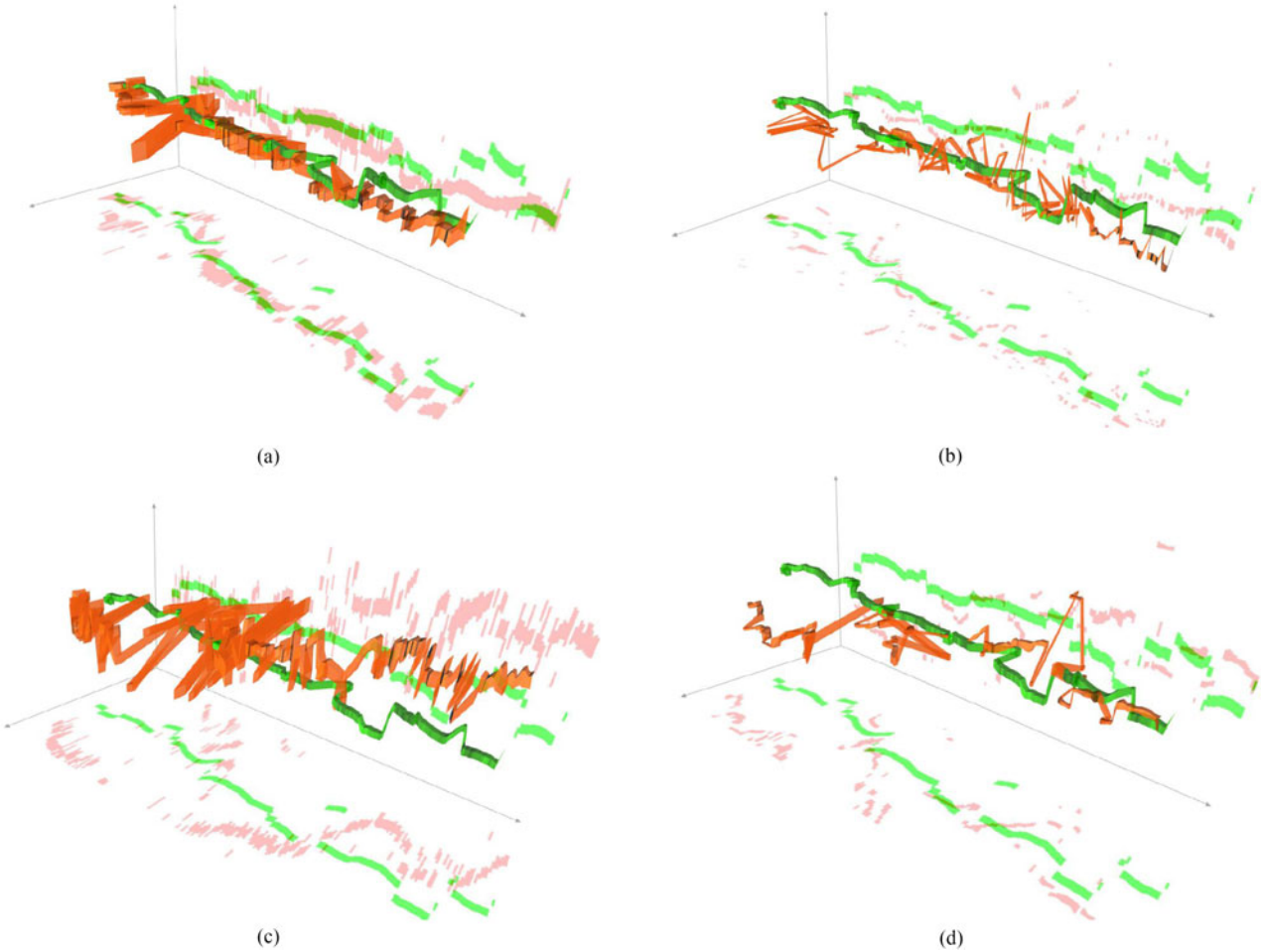


Fig. 9. Additional visualization of saliency prediction (orange) by saliency models compared with human's eye-tracking data (green) (a) AWS versus human. (b) SUN versus human. (c) SIG versus human. (d) SR versus human.

point. People's top-down knowledges in this case affect the behaviors a lot as they try to find attractive things based on the experience in daily life.

Bottom-up feature should be addressed properly: Fig. 8 shows two space-time cube views, with red cubes representing the saliency map predicted by saliency algorithm (GBVS) and the green cubes representing the human eye-tracking data. The size of cube area at each time frame encodes the size of AOI within the saliency map.

We can immediately find that the red and green cube overlap with each other within most of the time frames, which means the GBVS algorithm could predict salient area accurately and effectively in most video frames. Similar patterns are also found under the experiment of AWS algorithm [see Fig. 9(a)]. Moreover, comparing the space-time cubes side by side can help identify the potential common problems existed in the prediction algorithm. For example, the cubes in time segment [blue box on the left in Fig. 8(a)] reveals that both GBVS and AWS fail to predict the right salient area, while the area predicted by the two algorithm are almost the same. This pattern provides a visual hint for further examination of the video frames at that time segments and exploration of the reasons why the predicted ones do not match eye-tracking data.

In our evaluation, interestingly, we find that the two selected approaches for dynamic stimuli named RWRV and SR are beaten by traditional static saliency model, which shows the importance of bottom-up feature in saliency prediction. We pick out the time segments that the AOIs from human and computer do not overlap with each other for further analysis. The green and red boxes in Fig. 8 refer to AOIs from eye-tracking data and saliency algorithms respectively. In Fig. 8(a), human focus on the rider while the bottom-up saliency model regards the car as the most attractive thing. This suggests the necessity to detect high-level features such as human before extracting low-level features. This can also be observed in Fig. 8(b), where we can see the false detection of RWRV mainly lie in excessive emphasizing of low-level features such as color and boundary.

The size of salient regions should be limited: We qualitatively show more visualizations for the rest of the saliency models in Fig. 9. Except for AWS, the last three models barely matches the pattern people observe.

As mentioned above, the AOIs obtained from eye-tracking data are not only consistent in location, but also in size. Moreover, the size of region that is being fixated by human is not quite large generally, which suggests a further selection strategy after

TABLE III
USABILITY OF OUR VISUALIZATION DESIGN IN ANALYSIS TASKS

Tasks
1 .Explore the attractive features or AOI in a video.
2. Find the sequence of objects that have been focused.
3. Explore the duration of fixating on each object.
4. Gain a spatiotemporal view of the whole dynamic data.
5. Evaluate the synchrony of attention from human and computer.
6. Explore the weakness and strength of saliency models in prediction.
7. Explore the role of features from different levels in saliency prediction.

obtaining big areas containing salient features. For instance, a person is salient does not mean people will look at the whole body because the most attractive region is the face.

Our findings are valuable to be considered in building new saliency models and we think they are also helpful to be utilized for improving the existing saliency models.

C. User Feedback

Two experts (EA and EB) in visual saliency studies from two universities were asked to work on this study, identify research problems, and collect design requirements. The system was iteratively improved throughout the frequent meetings with the domain experts. The case studies were conducted when the system was ready. The experts provided interesting insights into the research findings. Their feedback is summarized as follows:

Visualization Design: The visual design of our system was received very well by both EA and EB. They agreed that the tool is engaging, and easy to use, and were very impressed by the interactive features. EA said that the 3-D view of eye-tracking data distribution is intuitive and helpful for data analysis and exploration. He mentioned that the data trace showing both human and computer model helps his exploration of attractive features and understanding of the way people observe as well as how it differs from computer algorithm. He also added that the idea to show the different traces of eye-tracking data of a whole video at once instead of watching and searching through each frame is wonderful and saves time. EB was impressed by the visualization view of our tool. He said that the design "allows me to easily switch to frames of interests to look for the factors that affect the pattern people and computer behave and lead to the difference. However, despite the experts appreciation of the overall design, they found the difference heat map difficult to understand because of the unaligned direction when compared with the data in 3-D view.

Usability: Both users confirmed the usefulness and effectiveness of the system and wanted to use the system in teaching and research. EA said that "The system is a great tool. I can use it to quickly find interesting frames that contain important features judged by both human and computer model". He especially liked the selectable 3-D trace, which allows him to select and see the exact image and saliency easily. EB noted that the system is not only useful for data analysis but also helpful for easily communicating their findings to colleagues or a wider audience. The usability in analysis tasks summarized by users are listed in Table III.

Limitation: Both users complained about the occlusions problem. They said that in our design, the opacity is too high, making it hard to observe from the occlusion. However, when opacity is set to lower than 0.6, the surfaces of traces are frequently mixed with each other which makes it hard to judge the original distribution of data. A more sophisticated way of visualization design for effective observation could be a possible avenue for future work.

Suggestion: The users provided valuable suggestions to improve the design. EA suggested the 2D projections be added to 3-D view of data to reduce the problems caused by 3-D occlusions. EB suggested that the design be kept as simple as possible so that common users can handle well to do researches in visual saliency. He also added that we show the corresponding frame beside or overlaid on difference heat map to make it easier for user to understand.

VII. CONCLUSION

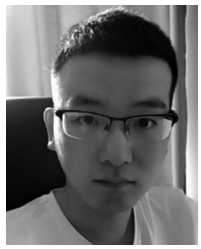
The existing method of visualizing eye-tracking data only consider human's behavior while in the field of building saliency detection model, a more important thing is to explore the different observation pattern between human and computer when being presented visual stimulus. In this paper, we propose visualization designs for the purpose of analyzing saliency models and eye-tracking data intuitively. We provide spatial and temporal views of visual attention data from both human and computer, which allows a wide collection of different analysis tasks. Lastly, we list our key findings from our analysis on human and computer's attention data to reveal the importance of stability and low-level feature, suggesting better design of saliency detection algorithm and showing deeper understanding of human observation pattern.

For future work, we plan to annotate the whole database we use in order to track objects for a more deeper analysis. Further, the findings from our analysis allow us to build saliency model that combines the key factors for better performance.

REFERENCES

[1] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.
[2] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*, New York, NY, USA: Springer-Verlag, 1987, pp. 115–141.
[3] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.
[4] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *J. Vis.*, vol. 9, no. 12, p. 10, 2009.
[5] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans., Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.
[6] A. Borji, D. N. Sihite, and L. Itti, "Computational modeling of top-down visual attention in interactive environments." in *Proc. BMVC*, 2011, pp. 1–12.
[7] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Proc. 2007 IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.
[8] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *SPIE*, vol. 5200, pp. 64–78, Jan. 2004. .
[9] A. Torralba, "Modeling global scene factors in attention," *JOSA A*, vol. 20, no. 7, pp. 1407–1418, 2003.

[10] A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Mach. Vis. Appl.*, vol. 22, no. 1, pp. 61–76, 2011.
[11] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 150–165, 2010.
[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[13] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 161–180, 2011.
[14] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inform. Process. Syst.*, 2005, pp. 155–162.
[15] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4, pp. 411–430, 2000.
[16] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–6.
[17] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.
[18] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.
[19] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vis.*, vol. 8, no. 14, pp. 341–360, 2008.
[20] A. Nuthmann and J. M. Henderson, "Object-based attentional selection in scene viewing," *J. Vis.*, vol. 10, no. 8, pp. 381–400, 2010.
[21] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep.-Oct. 2009, pp. 2106–2113.
[22] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 1–20, 2014.
[23] C. Shen and Q. Zhao, "Webpage saliency," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 33–46.
[24] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 17–32.
[25] W. Kienzle, F. A. Wichmann, M. O. Franz, and B. Schölkopf, "A nonparametric approach to bottom-up visual saliency," in *Proc. Adv. Neural Inform. Process. Syst.*, 2007, pp. 689–696.
[26] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 438–445.
[27] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2552–2564, Aug. 2015.
[28] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 15–15, 2009.
[29] G. Tien, M. S. Atkins, and B. Zheng, "Measuring gaze overlap on videos between multiple observers," in *Proc. Symp. Eye Tracking Res. Appl.* ACM, 2012, pp. 309–312.
[30] R. B. Goldstein, R. L. Woods, and E. Peli, "Where people look when watching movies: Do all viewers look at the same place?" *Comput. Biol. Med.*, vol. 37, no. 7, pp. 957–964, 2007.
[31] P. Marchant, D. Raybould, T. Renshaw, and R. Stevens, "Are you seeing what i'm seeing? An eye-tracking evaluation of dynamic scenes," *Digit. Creativity*, vol. 20, no. 3, pp. 153–163, 2009.
[32] T. Smith and J. Henderson, "Attentional synchrony in static and dynamic scenes," *J. Vis.*, vol. 8, no. 6, pp. 773–773, 2008.
[33] K. Holmqvist *et al.*, *Eye Tracking: A Comprehensive Guide to Methods and Measures*. New York, NY, USA: Oxford, 2011.
[34] A. A. Bojko, "Informative or misleading? Heatmaps deconstructed," in *Proc. Int. Conf. Human-Comput. Interact*. Springer, 2009, pp. 30–39.
[35] A. T. Duchowski, M. M. Price, M. Meyer, and P. Orero, "Aggregate gaze visualization with real-time heatmaps," in *Proc. Symp. Eye Tracking Res. Appl.*, ACM, 2012, pp. 13–20.
[36] D. S. Wooding, "Fixation maps: quantifying eye-movement traces," in *Proc. 2002 Symp. Eye Tracking Res. Appl.*, ACM, 2002, pp. 31–36.
[37] H. Y. Tsang, M. Tory, and C. Swindells, "Visualizing sequential fixation patterns," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 953–962, Nov./Dec. 2010.
[38] M. Raschke, X. Chen, and T. Ertl, "Parallel scan-path visualization," in *Proc. Symp. Eye Tracking Res. Appl.*, ACM, 2012, pp. 165–168.
[39] G. Andrienko, N. Andrienko, M. Burch, and D. Weiskopf, "Visual analytics methodology for eye movement studies," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2889–2898, Dec. 2012.
[40] P. Gatalsky, N. Andrienko, and G. Andrienko, "Interactive analysis of event data using space-time cube," in *Proc. 8th Int. Conf. Inform. Vis. IV*, 2004, pp. 145–152.
[41] M. Chen *et al.*, "Visual signatures in video visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 5, pp. 1093–1100, Sep./Oct. 2006.
[42] R. P. Botchen *et al.*, "Action-based multifield video visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 4, pp. 885–899, Jul./Aug. 2008.
[43] X. Li, A. Çöltekin, and M.-J. Kraak, "Visual exploration of eye movement data using the space-time-cube," in *Proc. Int. Conf. Geographic Inform. Sci.*, 2010, pp. 295–309.
[44] A. T. Duchowski and B. H. McCormick, "Gaze-contingent video resolution degradation," *Proc. SPIE*, vol. 3299, pp. 318–329, 1998.
[45] K. Kurzhals and D. Weiskopf, "Space-time visual analytics of eye-tracking data for dynamic stimuli," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2129–2138, Dec. 2013.
[46] J. C. Roberts, "State of the art: Coordinated & multiple views in exploratory visualization," in *Proc. 5th Int. Conf. Coordinated Multiple Views Explor. Vis., 2007*, 2007, pp. 61–71.
[47] Y. Sawahata *et al.*, "Determining comprehension and quality of tv programs using eye-gaze tracking," *Pattern Recog.*, vol. 41, no. 5, pp. 1610–1626, 2008.
[48] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cogn. Comput.*, vol. 3, no. 1, pp. 5–24, 2011.
[49] A. Santella and D. DeCarlo, "Robust clustering of eye movement recordings for quantification of visual interest," in *Proc. 2004 Symp. Eye Tracking Res. Appl.*, 2004, pp. 27–34.
[50] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenesshort title?" *J. Vis.*, vol. 14, no. 8, pp. 5–5, 2014.
[51] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image Vis. Comput.*, vol. 30, no. 1, pp. 51–64, 2012.
[52] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inform. Process. Syst.*, 2006, pp. 545–552.
[53] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 621–640, 2008.

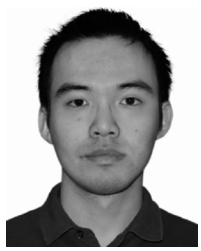


Haoran Liang received the B.Eng. degree in computer science from Zhejiang University of Technology, Hangzhou, China, in 2011, where he is currently is working toward the Ph.D. degree.

His research interests include computer vision, machine learning, and visualization.

Ronghua Liang (M'06) received the B.Sc. degree from Hangdian University, Hangzhou, China, in 1996, and the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2003.

He worked as a Research Fellow with the University of Bedfordshire, Bedfordshire, U.K., from April 2004 to July 2005, and as a Visiting Scholar at the University of California, Davis, CA, USA, from March 2010 to March 2011. He is currently a Professor of computer science and the Executive Dean of College of Information Engineering with Zhejiang University of Technology. His research interests include computer vision, information visualization, and medical visualization.



Guodao Sun received the B.Sc. degree in computer science and technology and the Ph.D. degree in control science and engineering from Zhejiang University of Technology, Hangzhou, China, in 2010 and 2015, respectively.

He is an Assistant Professor with the College of Information Engineering, Zhejiang University of Technology. His main research interests include urban visualization and visual analytics of social media.