The prediction of head and eye movement for 360 degree images

Article · May 2018

DOI: 10.1016/j.image.2018.05.010

CITATIONS
3

READS
68

3 authors:

Yucheng Zhu
Shanghai Jiao Tong University
14 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Guangtao Zhai
Shanghai Jiao Tong University
262 PUBLICATIONS **2,945** CITATIONS
SEE PROFILE

Xiongkuo Min
Shanghai Jiao Tong University
43 PUBLICATIONS **253** CITATIONS
SEE PROFILE

Signal Processing: Image Communication 69 (2018) 15–25

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

The prediction of head and eye movement for 360 degree images

Yucheng Zhu, Guangtao Zhai *, Xiongkuo Min

Insti. of Image Commu. & Infor. Proce., Shanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Keywords:
VR
Omnidirectional
360 degree
Panoramic
Saliency detection
Head motion
Head–eye motion
Scanpath

ABSTRACT

Estimating salient areas of visual stimuli which are liable to attract viewers' visual attention is a challenging task because of the high complexity of cognitive behaviors in the brain. Many researchers have been dedicated to this field and obtained many achievements. Some application areas, ranging from computer vision, computer graphics, to multimedia processing, can benefit from saliency detection, considering that the detected saliency has depicted the visual importance of different areas of the visual stimuli. As for the 360 degree visual stimuli, images and videos should record the whole scene in the 3D world, so the resolutions of panoramic images and videos are usually very high. However, when watching 360 degree stimuli, observers can only see part of the scene in the view port, which is presented to the eyes of the observers through the Head Mounted Display (HMD). So sending the whole video, or rendering the whole scene may result in the waste of resources. Thus if we can predict the current field of view, then focuses can be put to the streaming and rendering of the scene in the current field of view. Further more, if we can predict salient areas in the scene, then more fine processing can be done to the visually important areas. The prediction of salient regions for traditional images and videos have been extensively studied. However, conventional saliency prediction methods are not fully adequate for 360 degree contents, because 360 degree stimuli own some unique characteristics. Related study in this area is limited. In this paper, we study the problem of predicting head movement, head–eye motion, and scanpath of viewers when they are watching 360 degree images in the commodity HMDs. Three types of data are specifically analyzed. The first is the head movement data, which can be regarded as the movement of the view port. The second is the head–eye motion data which combines the motion of the head and the movement of the eye within the view port. The third is the scan-paths data of observers in the entire panorama which record the position information as well as the time information. And our model is designed to predict the saliency maps for the first two, and the scanpaths for the last one. Experimental results demonstrate the effectiveness of our model.

1. Introduction

Virtual reality (VR) technology appeared in last century. Consumers' enthusiasm and the extensive applications in military promoted people's exploration of VR technology. In recent years, with the advancements of wearable equipment and mobile Internet, VR technology has been developing rapidly. A market research reported that the global market of VR related products such as VR cameras, head mounted display (HMD) and VR contents will reach twenty-five billion US dollars by 2021 [1]. Many unique characteristics contribute to the popularity of VR, among which the most distinctive two features are the completely immersive experience and innovative interaction mode. More specifically, VR displays panoramic videos in a spherical canvas to simulate a virtual environment around viewers. When the user is moving, sensors equipped on the device will record the data of orientations and positions of head. As soon as the motion data is transmitted to the processor, the processor will change the displayed scene accordingly [2].

VR provides an immersive experience in virtual environment through interactions of body movements, and it has been used in many fields such as entertainment area, medical application, education, and so on. In order to ensure the quality of viewing experience at the user end, it is essential to use stereoscopic panoramas with high frame rate and high resolution. As a consequence, panoramic images will contain much more information than traditional images. Compared with traditional images and videos, panoramic contents own some unique characteristics, which call for specialized processing techniques. For example, some panoramic images reach the resolution of 18000 × 9000, and such a large image will make the storage and transmission of VR content difficult. Besides, there are two separate views for left and right eyes, which can be used to bring stereoscopic perception, but will increase the data size at the

* Corresponding author.
E-mail addresses: zyc420@sjtu.edu.cn (Y. Zhu), zhaiguangtao@sjtu.edu.cn (G. Zhai), minxiongkuo@sjtu.edu.cn (X. Min).

https://doi.org/10.1016/j.image.2018.05.010
Received 21 September 2017; Received in revised form 16 May 2018; Accepted 16 May 2018
Available online 22 May 2018

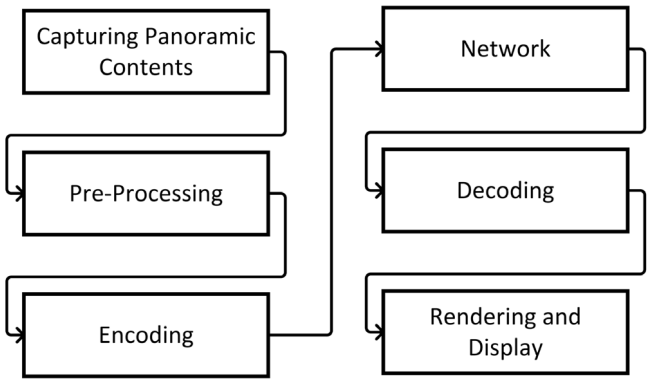


Fig. 1. Block diagram of a typical VR system, which consists of six main steps including capturing, pre-processing, encoding, transmission, decoding and display.

same time. For clearer illustration, Fig. 1 gives the block diagram of a typical VR system.

In Fig. 1, the whole chain begins with the capturing of panoramic contents. In this phase, multi-camera setup, such as cameras with dual lens and the synchronized multi-camera array, can be used to record the 360° scene. The synchronized camera array ensures precise spherical capture before and after each shot. Captured files from each sub-camera in the multi-camera setup will undergo pre-processes including color correction, stitching, etc. However, managing files is a constant challenge for capturing 360° content, because the files are huge. For the purpose of efficient storing and streaming, compression for the video is necessary to achieve a balance between file size and quality of content. After the compression in the encoding phase, panorama is transmitted through the network to the end user. The encoded bitstream will reach the consumer device at the user end, such as mobile phone and computer, and will be further decoded by the implemented decoder. In the final step, the decoded content will be rendered by local device prior to displaying. The process may include some operations like filtering, re-sampling, etc. After that, the image or video will be displayed in the commodity HMDs such as HTC Vive, Samsung Gear VR, and Oculus Riâft. When watching 360° stimuli, observers can only see part of the scene in the view port, which shows the views for each eyes as seen within the HMD. So sending the whole video, or rendering the whole scene may result in the waste of resources. Fig. 2 shows the image of large resolution and the limited viewing angle in the view port.

To save resources, we can predict the current field of view, and then we can focus on the streaming and rendering of the scene in the current field of view. Further more, we can predict salient areas in the scene for more fine processing of the visually important areas. The prediction of salient regions for traditional images and videos have been extensively studied. In the past 20 years, many saliency prediction models have been proposed for traditional 2D images [3]. However, conventional saliency prediction methods are not fully adequate for 360° contents, because 360° stimuli own some unique characteristics. The related research, including 360° content oriented saliency prediction, compression and quality assessment are just at the initial stage.

2. Our work

Considering that there is very few study in the field of panorama oriented saliency prediction, in this paper, we study the problem of predicting head movement, head–eye motion, and scanpath of viewers when they are watching 360° images in the commodity HMDs. There are three types of data that are under specifical analysis. The first is the head movement data, which can be regarded as the movement of the view port. The second is the head–eye motion data which combines the motion of the head and the movement of the eye within the view port. The third is the scan-paths data of observers in the entire panorama which records the position information as well as the time information. And our model is designed to predict the saliency maps for the first two, and the scanpaths for the last one.

Actually, many visual attention models have been proposed [4]. Bottom-up, top-down, or hybrid features are common in saliency detection models, and many features have been proven to be effective. Some other methods take the advantage of deep learning and achieve significant performance gain. We propose a multi-plane projection method and equator bias to simulate the viewing behavior of human eyes in HMDs. Visual attention will be detected after the projection and then back projected into the equirectangular format. By integrating with multi-plane projection and equator bias, 2D visual attention models can be somewhat effective to predict the visual attention in VR panoramas.

In this paper, we develop a framework to integrate features into the prediction of visual attention in VR panoramas. In our solution, we leverage and modify several attributes that might guide the deployment of attention. We employ the guiding attributes including color, orientation and spatial frequency. Besides, the compositional balance is one of agreed-upon principles of aesthetics. Motivated by experimental results, we assume that fixations satisfy the local compositional balance. So we take symmetry as another contributing feature. But only low-level features may not effectively extract some high-level features which are beneficial for prediction of visual attention. We leverage the Deformable Part Model to detect cars and pedestrians. Experimental results confirm that above mentioned features can effectively guide the attention in VR panoramas.

Eye movement is not the only user behavior. When exploring in the virtual environment, users will move eyes while turning heads. We think that the head movement is a reaction based on the visual stimulus, but an inert one when compared with movements of eye. Some properties including relative positions, isotropic trend and the stronger equator bias are observed and employed into the prediction of head movement in our solution. Experimental results demonstrate the effectiveness of our model.

In our solution to predict the scanpath, we determine the position of fixations by optimizing the local compositional balance. There forms a graph when we take the fixations as nodes and take the connections between fixations as edges. The transfer probability between fixations is proportional to the dissimilarity between areas and inverse proportional to the distance. The scanpath is determined by maximizing the sum of transfer probability on the path.

3. Related work

In this section, we shortly review fields closely related to this study, including the projection methods of 360° contents, saliency prediction, and quality assessment. These areas address different key problems in the whole processing chain of omnidirectional content. The direct problem need to be studied is the projection method. Some projection methods are introduced for the benefit of content coding and transmission, while some others are designed to optimize the representation in the viewing of human eyes. In this paper, we design a multi-view projection to simulate the views in HMDs. As for the quality assessment [5–7], it is one of the most direct application fields of saliency prediction [8], for example, the saliency-based quality assessment for omnidirectional images. Measuring the quality of panoramic image/video with conventional methods does not represent the quality fairly. In order to enhance the performance, the predicted salient regions can be employed into the pooling stage. The reviews of related papers can be found below.

3.1. Projection method

Panoramic contents can be represented by a sphere in 3D spherical coordinates, but the usual practice is to map the panorama from 3D coordinates into 2D coordinates. Because there have been many studies

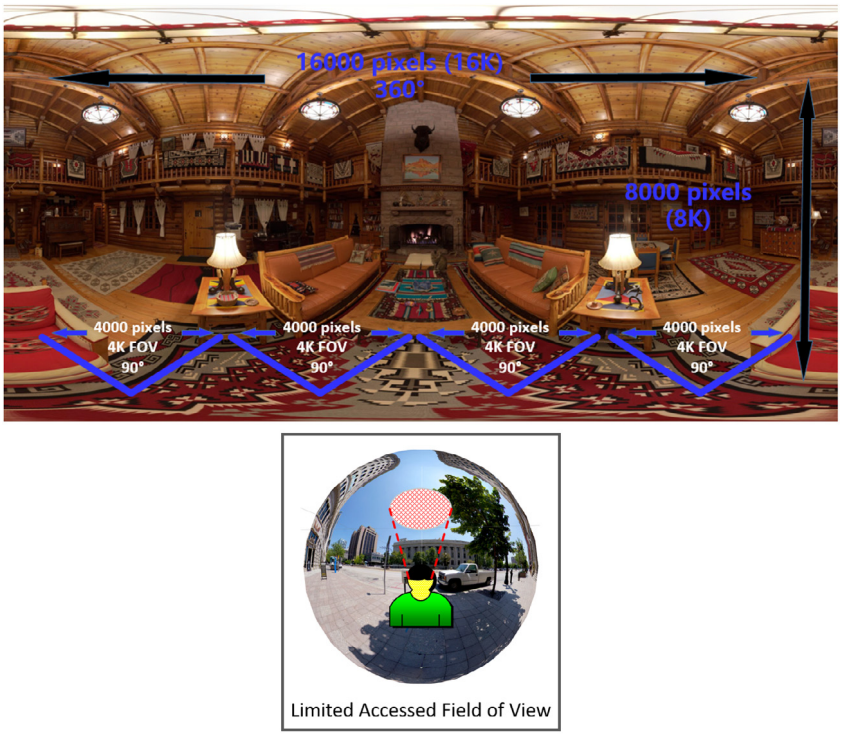


Fig. 2. One of the most important difference between conventional 2D image and 360 image is that, 360 degree images have much larger resolution and file size to provide the user a larger field of view. But with the HMD, user can only get to see a small part of the whole view. Though user can change their orientations and positions to explore other field of view, the instantaneous accessed field of view is limited.

for conventional 2D contents that can be used for reference in the fields of video coding, saliency prediction, quality assessment, rendering, *etc.* Various projection methods can be used, *e.g.*, cylindrical [9], cubic [10], *etc.* The projection is necessary in many processes for panorama, for example, coding of panorama using different mapping methods have been studied in related literatures [11–15].

Among the projection methods, the equirectangular projection is the most widely used method in related applications due to its rectangular projection plane, which is widely supported in software development environments. However, this mapping method stretches some areas in the 2D coordinates to ensure a rectangular image after projection and thus contains a relatively great amount of redundant data. By contrast, the pseudo-cylindrical projections will not produce redundant information in polar areas, but the projection map is non-rectangular. Pseudo-cylindrical projections for navigation, rendering and interaction purposes in VR applications have been studied in literatures [16–18]. And the comparisons of the two projection methods can be found in [19].

Some other researches designed the mapping methods to optimize the representation in the viewing of human eyes. Carroll et al. [20] proposed a new projection method that would adapt to the content in the scene. The projection method exploited the fact that human visual system tends to focus on some specific areas (salient areas) by employing a local adaptively varying projection to preserve salient areas. Zelnik et al. [21] proposed a multi-plane perspective projection method that incorporated multiple local projections into the same panorama. The parameters of projection method locally adapted to the content of the images in order to reduce distortions introduced by the projection and produce more compelling results.

3.2. Saliency prediction

Some applications ranging from computer graphics, computer vision to the media encoding, transmission and quality assessment benefit from the saliency detection of visual stimuli in the aspect of providing the better quality of experience.

3.2.1. Subjective experiments

To understand visual attention in omnidirectional images, ground truth data must be gathered through subjective tests. Rai et al. [22] provided a freely available dataset of omnidirectional images together with results of head and eye movements obtained from a subjective experiment. The dataset contained sixty 360° images. The total number of observers that participated in the test was sixty-three which resulted in an average of 40–42 observers per stimuli. Head and head–eye saliency maps together with fixation data and head motion data were calculated from eye-tracking information and HMD sensor information. Lo et al. [23] created a 360° video viewing dataset in head-mounted virtual reality. There were ten testing videos and fifty subjects in the subjective experiment. The content-related information including image saliency maps were generated by using Convolutional Neural Network, while the motion maps were generated by analyzing the optical flow. And the sensor-related information including head orientations and positions were obtained from HMD sensors.

3.2.2. Objective models

Saliency maps can also be predicted using models of visual attention. In [24], two types of features including the sensor-related features and content-related features, and Recurrent Neural Network were employed to predict fixations in 360° videos. HMD orientations were obtained by sensors and were aligned to 360 degree videos to identify where the viewer was watching, while the image saliency maps and motion maps were calculated as content-related features. Abreu et al. [25] created a database for omnidirectional images (ODIs). There were twenty-one ODIs in the dataset which were watched by thirty-two viewers. The trajectories of viewport center were collected and transformed into saliency maps. Besides, a post-processing method was proposed on the basis of equator bias tendency to fast design ODIs oriented methods from current saliency models. In [26], Rai studied some saliency weighting strategies for generating saliency maps from head movement data which can be used as spatial pooling methods for the quality assessment of omnidirectional images. In [27], Hu proposed an online human-like

agent to automatically navigate a panoramic video. The underlying principles were to capture most interesting events and then make smooth transitions for viewers. The proposed model only focused on foreground objects and extracted the object-based observations. Then the trained selector block took the observations as input and selected the main object. Finally, given the location and motions of main object, the regressor block regressed the viewing angle. In [28], Sitzmann explored the viewing behavior and visual attention in VR. To assess whether watching conditions, scene contents and starting points have an effect on viewing behavior, they conducted a subjective experiment. 22 panoramas including indoor and outdoor scenarios were used as stimuli. For each scene, four different starting points and three different viewing conditions were tested. As a result, they recorded 1980 head and gaze trajectories from 169 people. The analysis of the data and other applications were also provided.

3.3. Quality assessment

As with any multimedia content, each step in the VR system may introduce certain distortions degrading the quality perceived by the end user, so the quality assessment is important to the development of 360° content technology. However, only a few studies have been published addressing this issue.

3.3.1. Subjective experiments

Some experimental conclusions and ground truth data can be obtained through subjective experiments. Evgeniy et al. [29] created a testbed for subjective quality evaluation of omnidirectional visual content. There were thirty images in the database among which five images were for training and twenty-five images were for testing. Images were generated using JPEG encoders of four different quality factors. Total 48 subjects participated in the subjective assessment. Subjective scores, tracks of viewing directions and the spending time of subjects on each stimulus were provided. Correlation analysis of objective metrics and subjective scores were also listed. Duan et al. [30] presented an Immersive Video Quality Assessment Database for the video quality assessment in virtual environment. The database contained ten raw videos and each lasted for fifteen seconds. To simulate quality degradations, three resolutions were set. And under every resolution, different bit rates and frame rates were set to simulate different bandwidth requirements.

3.3.2. Objective metrics

Some models were designed to evaluate quality objectively. In [31], both low-level vision factors and high-level vision factors were considered. The quality assessment metric analyzed three main video artifacts: jerkiness, blockiness and blur. Besides, regions near the seam between adjacent cameras and areas where the motion occurs were likely to attract viewers' attention. And larger weights were assigned to these areas. Zakharchenko et al. [19] discussed the immersive content distribution format and the perceptual quality estimation. To estimate the difference between original and reconstructed images, the authors proposed two methods. The first was to remap image to Craster parabolic projection. The second was to consider the anisotropy of different projection methods and introduced anisotropic weights to simulate differences of isotropy between areas. Yu et al. [32] compared the peak signal-to-noise ratio (PSNR) between viewports that were generated according to the head motion data that were recorded by HMDs. To reduce the dependence of recorded data, the saliency weighting over possible viewing directions was measured. The authors further discussed the impact of different panoramic map methods on the perceptual quality and coding efficiency such as equirectangular, Lambert cylindrical equal-area, dyadic, and cubic projections.

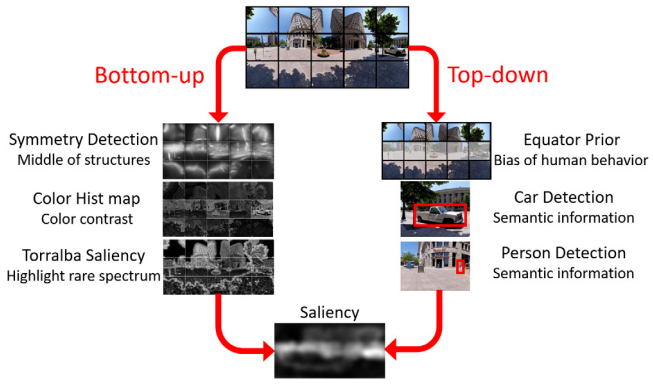


Fig. 3. The framework of our saliency prediction method. Images in equirectangular format will be first projected into several blocks to simulate view ports in HMDs. Then bottom-up and top-down features are extracted in each block. Feature maps are fused and further mapped into the overall saliency map which is in equirectangular format.

4. Our models

In this section, we propose a model to predict head movements, head–eye motions, and scanpaths of viewers when watching 360° images in the commodity HMDs. By predicting the movement of viewers' head, we can get the center trajectories of view ports, which can be regarded as approximate shifts between regions of interest. To obtain precise salient areas, eye motions should be combined with head movements. As for scanpath, the position of fixations as well as the duration of fixations should be predicted to determine the shift of fixations over time. In our model, we design a new projection method and employ equator bias to mimic the viewers' behavior in HMDs. Our model, which includes the top-down and bottom-up features, is designed to predict the saliency maps for the first two, and the scanpaths for the last one, and is proven to be effective. The diagram in Fig. 3 illustrates the framework of our saliency prediction method. The omnidirectional image will be firstly projected into multi-view blocks. Bottom-up and top-down features for each block are extracted and fused. The saliency maps for all blocks are finally mapped into the overall saliency map in the equirectangular format.

4.1. Our projection method

A multi-plane projection method is proposed to simulate the viewing behavior of human eyes in HMDs. We set some planes tangent to the sphere to cover the field of view in VR panoramas. As a result a small field of view on the sphere will be projected onto the plane. To make the field of view covered by each plane more like the view port in HMDs. We research the popular commodity HMDs and find that the viewing angles of most HMDs are around 100° in both horizontal and vertical directions. When users are watching panoramic stimuli, the transitions between view ports are smooth. By controlling the rotation angle between adjacent planes, we can adopt a small rotation angle to model the smooth transition.

As illustrated in Fig. 4, the image in equirectangular format will be mapped to sphere in 3D coordinates. A rectangular view port of fixed width and height, whose center is located on the equator, is initialized in 3D coordinates as the initial projection plane. The content within the visual field of the view port on the sphere will be projected onto the plane. Then the view port will be rotated, and the center of the plane will be aligned to points of different latitudes and longitudes on the sphere. As a result, a series of view-port blocks are derived to simulate the views of human eyes in HMDs. Before the prediction of salient regions, the omnidirectional image will be first mapped into multi-view blocks.

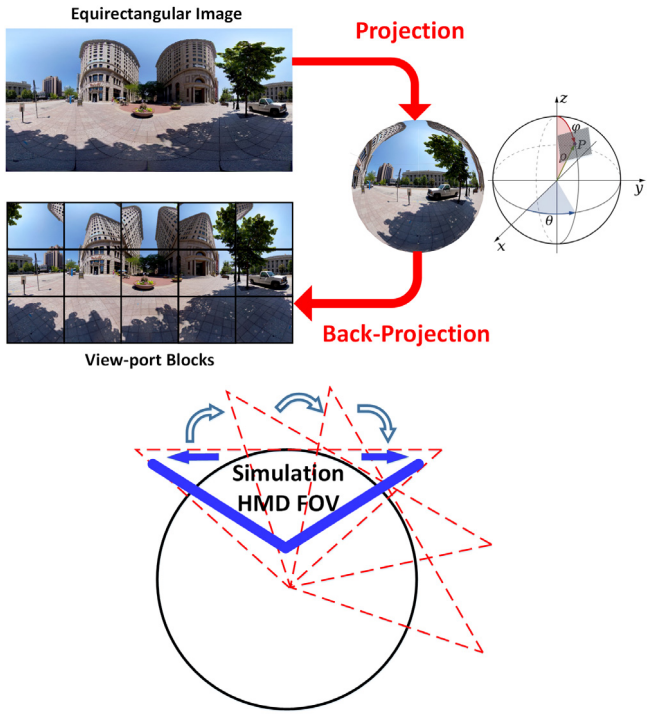


Fig. 4. The multi-view projection. The image in equirectangular format will be mapped to sphere in 3D coordinates. And then the multi-view projection is employed to project the sphere into view-port blocks which are tangent to the sphere in different directions.

4.2. Equator bias

According to experimental results from literatures [25,28], there is a clear equator bias when viewers are watching in HMDs. Statistics show that viewers are more likely to view content in the surroundings of equator rather than areas near the poles. Such a pattern of viewing behavior is expected in the common case, as the erect position is a natural head position. The probability of the situation when head is held erect is higher than that when head is tilted forwards or backwards. From another aspect, when taking photos, photographers tend to place prominent objects in the center areas. As for panoramic images, objects of interest will appear more frequently in the proximities of equator after the stitching.

4.3. Features to predict salient regions

We choose low-level features as they have already been proven to have underlying biological validity which contributes to the selectivity of human visual system (HVS). Statistics in the spatial domain and frequency domain are widely employed [33–35], and the color is also supported as one of the main attributes that guide the deployment of attention [36–38]. In our solution, the spatial frequency, orientation [34] and color [39] are based on some former designed features. We do the modification by integrating features with the contrast sensitivity and multichannel of HVS. The employed low-level features are listed below:

1. Statistics of different levels and orientations in the sub-band pyramid domain.
2. Statistics of different color channels in the spatial domain.
3. Estimation of compositional balance in the spatial domain.

The bottom-up selection of visual attention correlates with involuntary attention. It is an involuntary, fast and stimulus-driven mechanism [40]. The characteristics of bottom-up selection are attribute to properties of HVS. HVS serves as the passive selector, acknowledging some stimuli but rejecting others. The sensibility of our eyes to the spatial frequency and contrast has been widely researched [41,42]. It has been shown that the contrast sensitivity is a psychophysical phenomenal attributed to different properties of HVS. There is a minimum value of contrast, which is called the contrast detection threshold, and contrast sensitivity is the inverse of the threshold. When the value of contrast is below the threshold, the visual target is undetectable. It has been shown that the spatial frequency of the visual target has an impact on the threshold. The so-called contrast sensitivity function (CSF) has the spatial frequency as the independent variable and the contrast sensitivity as the dependent variable. The threshold changes with the spatial frequency. Considering the resulting profile of CSF, we assign different weights for contents of different spatial frequency on the image. Besides, experimental results have demonstrated the multichannel model of human vision [41]. HVS conducts a spatial-frequency decomposition of a stimulus locally in which the multiple spatial-frequency channels independently detect different frequency components. So images can be described and processed as a superposition of different spatial-frequency components.

Considering the contrast sensitivity and the multichannel of human vision. We construct a steerable pyramid on the achromatic component of panoramic images. Spatial filters of different orientations and frequency bandwidths are used for constructing levels of a steerable pyramid. After the formation of image pyramid, we calculate the histogram of components of different spatial frequencies and orientations to estimate the probability density function (PDF). Eq. (1) below computes the feature value of pixel s by the linear combination of different scales and orientations.

$$f_1(s) = \sum_{\forall k \in W} \alpha_k \log P_k^{-1}(I_s) \tag{1}$$

where α_k includes the weighted addition across orientations and scales. Four orientations including 2 in vertical direction and 2 in horizontal direction are given the same weight, while the weights between different scales are determined according to the profile of CSF. P_k sets the probability of intensity in subband k of the pyramid W. After the linear combination in the pyramid, we get the value at pixel s.

Color is supported as one of the attributes that guide the deployment of visual attention by large amount of convincing data [43]. A plausible way to use the color information is proposed. It includes the calculation of distribution and feature map for each of the RGB channels and the merging of feature maps for three channels. The feature map can be computed as Eq. (2) below.

$$O(s) = \sum_{\forall c \in C^{RGB}} \lambda_c \log P_c^{-1}(I_s) \tag{2}$$

where λ_c are the weights for color channels which are learned from the conversion of RGB to luminance value in a given color format (YUV). P_c sets the probability of intensity in different color channels. There seems to be plausible to enhance the saliency of a structure that is surrounded by a high color contrast. As the Eq. (3) shows, we multiply the color distance in the L*a*b* space with the weights in the spatial domain based on the distance between pixels.

$$f_2(s) = \frac{1}{k_s} \sum_{\forall s' \in \Omega} g_d(s' - s) C(I_{s'} - I_s) O(s) \tag{3}$$

where k_s is a normalization term:

$$k_s = \sum_{\forall s' \in \Omega} g_d(s' - s) \tag{4}$$

C measures the color distance in the L*a*b* color space and can be computed as Eq. (5) below.

$$C = [(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2]^{\frac{1}{2}}. \tag{5}$$

The function g_d sets the weight in the spatial domain based on the distance between the pixels. We use a Gaussian for g_d in the spatial

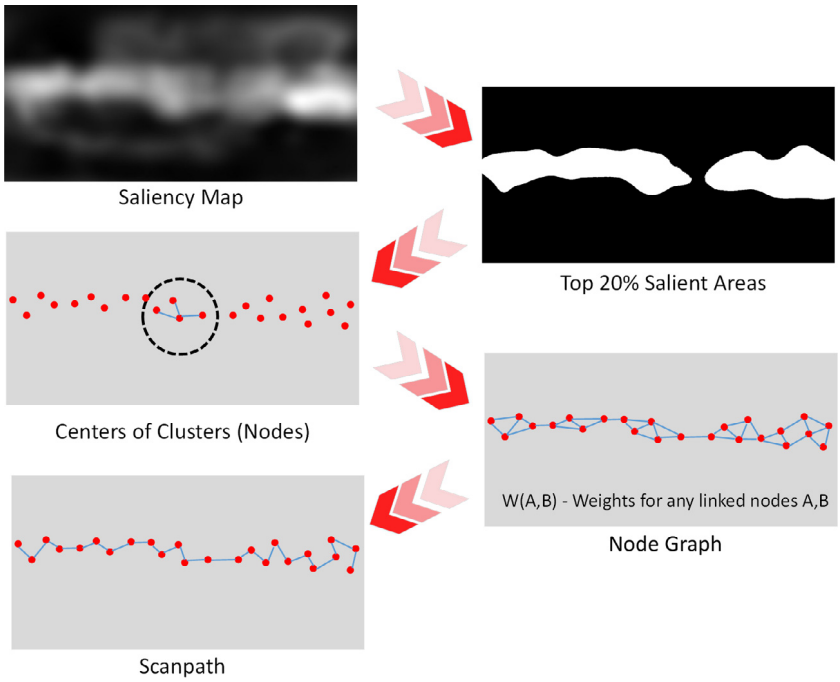


Fig. 5. The diagram of scanpath generation. We first binarize the saliency map by segmenting the top-salient areas. Then we conduct clustering for the binarized saliency map, and get the centers of the clusters. For each node, we link it to the near nodes within a certain range to form a node graph. Finally, scanpath can be generated by maximize the overall weights of the all path.

domain and control the width by the standard deviation parameter σ. Therefore, the value at a pixel s is enhanced mainly by significant local color contrast.

An unbalanced scene can cause discomfort to viewers, so unless this is our desired result, we usually set the composition to ensure that our images are balanced. For the image, the unit of measure is visual interest, that is, to achieve visual equilibrium, our image should be balanced by visual interest. The compositional balance was employed in seam carving, composition optimization, image attention retargeting, automated layout design, and image quality assessment [44–48]. Visual balance can be achieved in various ways, for example on horizontal and vertical axes. Symmetrical balance is the simplest type of balance to achieve. Because the panoramic image recorded the full surroundings, the composition rules could not be required at the time of shooting. But the composition rules can be used as reference when viewers alter the position of view port when they are watching the panoramic images. People tend to balance the composition of the scene in his view. Experimental results show that when subjects are asked to crop photographs the resulting center of the visual mass is nearby the original symmetry axis [49]. The visual draw that an element has is often referred to as its visual mass. Motivated by these findings, we enhance the local symmetry axes in the image. A low-level feature extraction algorithm [50] is included to detect symmetry axes in images. The detected symmetry maps are taken as f_3.

In addition to low-level features, high-level features are included into our model as well. Experimental results [39] showed that people tended to fix on some objects like people, car, faces, *etc.* Felzenszwalb's object detector [51] is used to extract our high-level features. The high-level features are listed below:

1. Equator bias due to tendency of human behavior.
2. Car and person detectors implemented by Deformable Part Model.

The diagram in Fig. 3 illustrates the framework of our saliency prediction method. The omnidirectional image will be firstly projected into multi-view blocks. Bottom-up and top-down features for each block are extracted and fused. The saliency maps for all blocks are finally mapped into the overall saliency map in the equirectangular format. Table 1 shows experimentally that each feature adds accuracy by combining different features and demonstrates that every feature makes a contribution to the prediction. As for the projection and equator bias, experimental results show that it plays an important role in the model.

4.4. From head motions to head plus eye movement

When watching panoramic images, the viewing patterns of observers are combinations of movements of head and eye. Experimental results [26] showed that in the view port of HMDs, the locations that viewers usually look at are the areas offset from the center of view ports for about twelve degrees. There is an inference that the offset is the antecedent movement of eye that precedes the head. Eyes will first move in a certain direction then follows the head, but single eye movement is sufficient when the exploration is near the boundary of salient region. So the area of head movement should be encircled by salient areas. Besides, there was an isotropic trend of the viewing directions and frequencies in the view port [26]. The encirclement and isotropic make it reasonable to assume that head motions are oriented at the centers of local salient areas. Although equirectangular projection stretches in the horizontal direction, it stretches equally and will not change the symmetry. Subjective experiments can also reveal that there is a strong equator bias for head motions which can be seen in Fig. 8. Fig. 8 presents the head motion related maps. From these images, there can be seen a strong equator bias for head motion. The top 20% maps show that the distributions of salient areas for head–eye motions are somewhat disperse, but the active areas for head motions are almost centered around equator. Based on the above observations and assumptions, we conduct the prediction of head motions in equirectangular format by removing the areas on boundaries, applying a stronger equator bias and highlighting top percent areas in the head–eye saliency maps.

4.5. Scanpath

Eye-tracking data can be classified into synchronic indicators and diachronic indicators [52]. The former focuses on events occur at

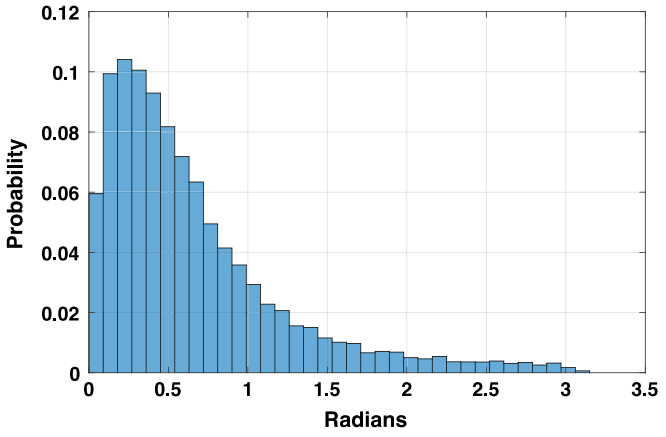


Fig. 6. The normalized histogram for rotation angles.

specific points of time, while the latter emphasizes events that occur over time. Among the diachronic indicators, the scanpath plays an important role in recording the shift of visual perception and attention over time. Scanpath data contains the fixation time and the position information of fixations. By recording the scanpaths, the direction and duration of the transfer from one fixation to another can be calculated, which is so called saccades of human eyes. A saccade is a very fast jump from one eye position to another. The movement is very fast, and the start of the latter fixation can be regarded as the end of the previous one.

When the composition of image is unbalanced, the composition may create tension and raise uneasy, disquieting responses in the viewer. The unit of measure is visual interest and people tend to balance the composition of scene in the field of view [49]. When watching panorama, part of the spectacle is rendered in the viewport and viewers can actively balance the visible scene with fast eye movements. The visual draw that an element has is often referred to as its visual mass, where this mass-analog takes into account both the area and the degree of saliency of visually salient regions [46]. When the fixations of viewers are located at the center of visual mass, the plausible visual equilibrium will be achieved.

Taking into account the compositional rules and its relationship with human preference, we extract local center of visual mass as the candidate fixations. To be detailed, we extract the top-salient areas in the conspicuity map and take them as the region of interest (ROI). Then we construct clusters based on the 2D location information of ROI. Here, the Lloyd's algorithm [53] is used for clustering, and we calculate centroid locations of each clusters. These centroids can be regarded as the local center of visual mass and the candidate fixations.

There shows a diagram for the prediction of scanpath in Fig. 5. There forms a graph when we take the fixations as nodes and take the connections between fixations as edges. The connection between nodes is that one node is restricted to only connecting with the neighboring nodes. And we assume that when watching panorama viewers will not go back to the part of the image for a second observation. To be more strict, we set that each node should be visited and can only be accessed once. We assume the transfer probability between fixations is proportional to the dissimilarity between areas [54] and inverse proportional to the distance. The free energy [55] theory reveals that the gap between real scene and brain's prediction causes people's attention. We assume that before the eye movement, the peripheral vision needs brain's prediction, and the center vision is the real scene. We take the dissimilarity between two areas as a measure of the gap. Dissimilarity between two nodes on the edge is measured by calculating the PSNR between local block patches. PSNR between two local blocks are calculated as Eq. (6) shows:

$$L(A,B) = PSNR(S_A, S_B) \tag{6}$$

where S is the block centered at the fixation node A, B. And we can compute the transfer probability as the Eq. (7) below.

$$W(A,B) \triangleq D(A,B) \cdot L(A,B) \tag{7}$$

where $D(A,B)$ measures the turning angle on the sphere.

The decision rule for the switch from one node to another is to maximize the sum of weights on the scanpath. Under the two constraints, it becomes a kind of Travel Salesman Problem (TSP) which is a NP problem and can be solved by the greedy algorithm. The formed graph is undirected and incomplete. Each node in the graph should have exactly one incoming edge and one outgoing edge. In our solution, we derive the local optimal solution by restricting the domain within five nodes. The suboptimal solutions are considered to have lower priorities. By combining local solutions of different priority levels we get different scanpaths. The priority of the path is calculated by averaging the priority of local solutions, and we select some paths of high priority as candidates. Through correlation analysis, we find that the duration of each fixation is independent with the saliency value and dependent with the number of fixations. We also calculate the rotation angles from one fixation to the next. The normalized histogram is shown in Fig. 6. We can see that most rotation angles are less than one radian, and the distribution peaks around 0.3 radian which can be used to modify the prediction of scanpath.

5. Experiments and results

In this section, we evaluate our model on the database for omnidirectional images [22]. They also provided a toolbox [56] which can be used to measure the performance of objective models for VR panorama. Several experiments are conducted to validate the effectiveness of features and the overall model.

5.1. Ground truth data

Rai et al. [22] provided a freely available dataset of omnidirectional images together with results of head and eye movements obtained from a subjective experiment. The dataset contained sixty original 360° images. The total number of observers that participated in the test was sixty-three which resulted in an average of 40–42 observers per stimuli. The raw eye movement and head motion data was provided including the gaze points for both eyes, the head rotation and translation parameters. Besides, the processed head saliency maps, head–eye saliency maps and scanpaths were calculated and provided. All the original images and processed saliency maps were in equirectangular format.

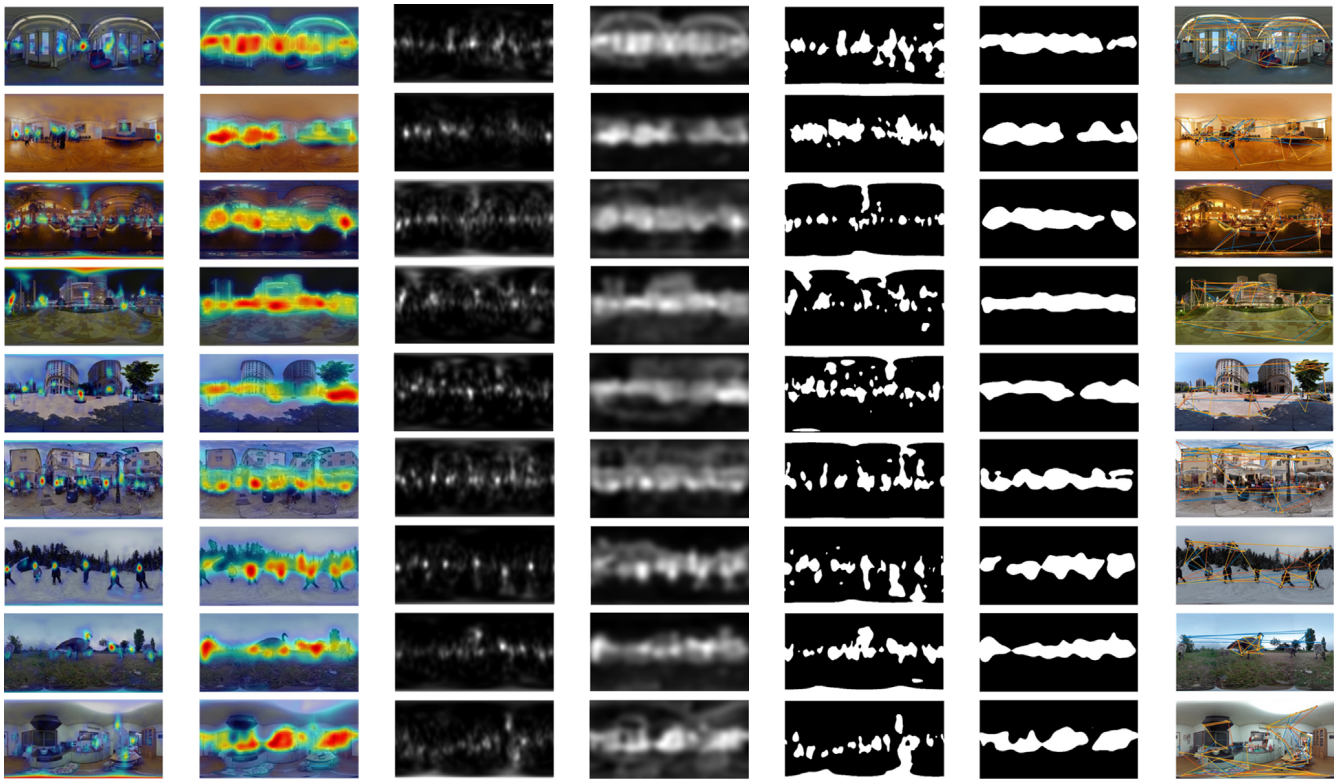
5.2. Evaluation metrics

Uniform sampling [57] is first conducted to eliminate the stretch around poles caused by the non-uniform characteristic of the equirectangular format. At the uniformly sampled points, the related 2D saliency evaluation metrics are used. The Normalized Scanpath Saliency (**NSS**) metric [58] quantifies the saliency values at the locations of eye fixations and use the variance of saliency map to normalize the results:

$$NSS = \frac{1}{N}\sum_{f=1}^{N}\frac{M(f)-\mu_M}{\sigma_M} \tag{8}$$

where f is the location of fixations, N is the total number of fixations, M is the saliency map, μ and σ are mean and standard deviation respectively.

The Kullback–Leibler (**KL**) divergence is used to estimate the dissimilarity between two distributions. KL metric was also employed to compare eye fixations with saliency maps [59]. The distribution of



(a) Heat map overlay Ground truth. (b) Heat map overlay Predicted. (c) Head–eye Saliency Ground truth. (d) Head–eye Saliency Predicted. (e) Head–eye top 20% Ground truth. (f) Head–eye Top 20% Predicted. (g) Scanpath Predicted.

Fig. 7. Experimental results under different scenes. (a) and (b) are original images overlayed with heat maps, (c) and (d) are head–eye saliency maps, (e) and (f) are top percent areas, (g) shows the predicted scanpath.

saliency map M is used to estimate the distribution of eye fixations map F.

$$KL = \sum_{v=1}^{X} F(v) \log\left(\frac{F(v)}{M(v)+\varepsilon} + \varepsilon\right) \tag{9}$$

where X is number of points to be compared, M and F are normalized distributions, and ε is used to avoid the division and log by zero.

The linear correlation coefficient (**CC**) metric measures the linear relationship between two variables

$$CC = \frac{corr(M, F)}{\sigma_M \cdot \sigma_F} \tag{10}$$

where $corr$ computes the correlation coefficient between M and F.

The metric **AUC-Judd** [60] is also used to compare between the saliency maps M and eye fixations. Pixels are extracted form saliency maps and the number of extracted points equals to that of fixations. By setting different thresholds, we can get the True Positives (TP) and False Positive (FP). Then the ROC curve can be drawn and the Areas Under Curve (AUC) can be computed. To eliminate the influence of stretch around the pole areas and make the evaluation reasonable and fair, an uniform sampling [57] is conducted on the sphere to refine the saliency map M and eye fixation map F.

5.3. Performance of the saliency model

In our model, we extract four different features including rare spectrum in subband domain, areas of high color contrast, symmetry map and object detections. Besides, we also assume that there exists an equator bias when viewers are watching in HMDs. And we conduct an experiment to analyze the contribution of these features:

(C1) using the rare spectrum only;

Table 1
The experimental results for head–eye motion prediction under cases of different feature types.

Feature type	CC	ROC	NSS	KL
C1	0.369	0.633	0.500	0.678
C2	0.271	0.619	0.439	0.722
C3	0.400	0.668	0.641	0.667
C4	0.512	0.715	0.902	0.512
C5	0.523	0.722	0.911	0.502
C6	0.435	0.689	0.718	0.661
C7	0.532	0.735	0.918	0.481

Table 2
The experimental results for head motion prediction.

Metrics	C1	C2	C3	C4	C5	C6	C7
CC	0.393	0.302	0.458	0.652	0.658	0.531	0.668
KL	1.163	1.211	1.126	0.674	0.663	0.998	0.654

Table 3
The experimental results for the combination of projection, equator bias and conventional models.

Models	CC	ROC	NSS	KL
GBVS	0.295	0.676	0.650	0.727
P-GBVS	0.383	0.691	0.717	0.635
Judd	0.392	0.576	0.650	0.727
P-Judd	0.465	0.687	0.824	0.563
MLNet	0.513	0.705	1.128	0.922
P-MLNet	0.521	0.723	1.235	0.868

(C2) using the color contrast only;

(C3) using the symmetry prediction only;

(a) Original image. (b) Head–eye saliency Ground truth. (c) Head saliency map Ground truth. (d) Head–eye top 20% Ground truth. (e) Head top 20% map Ground truth. (f) Head–eye saliency Predicted. (g) Head saliency map Predicted.

Fig. 8. Comparisons between head–eye saliency maps and head saliency maps. (a) are the original images, (b) and (f) are head–eye saliency maps, (c) and (g) are head saliency maps, (d) head–eye top 20% maps, (e) head top 20% maps.

Table 4
The grand challenge related results for head movement prediction.

Metrics	ZJU [63]	WHU [64]	SJTU ours	TUM_1 [65]	TUM_3 [65]	Roma Tre [66]	NWPU [67]
KL	0.444	0.515	0.654	0.745	0.737	0.810	1.137
CC	0.692	0.715	0.668	0.620	0.600	0.525	0.572

Table 5
The grand challenge related results for head–eye movement prediction.

Metrics	TUM_3 [65]	TUM_6 [65]	SJTU ours	WHU [64]	TUM_1 [65]	ZJU [63]	Trinity college [68]
KL	0.449	0.421	0.481	0.508	0.501	0.698	0.487
CC	0.579	0.615	0.532	0.538	0.554	0.527	0.536
NSS	0.805	0.807	0.918	0.936	0.915	0.851	0.757
ROC	0.726	0.722	0.734	0.736	0.747	0.714	0.702

(C4) using the symmetry prediction and object detections;

(C5) using the symmetry prediction, the object detections and rare spectrum;

(C6) using the symmetry prediction, the object detections, rare spectrum and color contrast but no equator bias;

(C7) using the symmetry prediction, the object detections, rare spectrum and color contrast with equator bias.

Experiments are conducted under cases of single feature type and combinations of features. The results are shown in Tables 1 and 2. From the results we can see that under the single-feature case, the symmetry detection achieves the best performance. And by combining different features, the results will be further improved, which demonstrates that every feature makes a contribution to the prediction. As for the projection and equator bias, experimental results show that it plays an important role in the model. To further prove its validity, we combine the projection and equator bias with other models including GBVS [61] and Judd [39]. Table 3 reports the results, from which we can see that there is an approximate gain of 20% by adding the projection method and the equator bias to GBVS (P-GBVS) and Judd (P-Judd). We further train the ML-Net model [62] on the training set and validate its performance on the test set. The data in equirectangular format is used for training at the first time. At the second time, we project images of equirectangular format into sub-blocks and integrate the equator bias into the model. We can see the good performance of training-based model and the enhancement of the refined version.

Fig. 7 presents the obtained heat maps, saliency maps and top percent maps. From the listed images, we can see that people and cars in the scene are salient objects, and some salient areas are of high contrast and symmetry. Fig. 8 presents the head motion related maps. From these images, we can see that there is a strong equator bias for head motion. The top 20% maps show that the distributions of salient areas for head–eye motions are somewhat disperse, but the active areas for head motions are almost centered around equator.

The ICME grand challenge related results are listed in Tables 4 and 5. Submitted models are evaluated by toolbox [56] on the private test database. The evaluation results confirm that our model gets the third place in the head movement prediction and the third place in the head–eye motion prediction. We give experimental results for some models. In the comparison of head movement prediction, results for models including ZJU [63], WHU [64], TUM [65], Rome Tre [66] and NWPU [67] are listed. In the comparison of head–eye movement prediction, results for models including TUM [65], WHU [64], ZJU [63] and Trinity College [68] are listed.

Table 6
The grand challenge related results for scanpath prediction.

Metric	Insight center for data analytics [69]	SJTU ours	WHU [64]
Vec	2.870	4.657	5.952

Table 7
Length of scanpath and standard deviation of duration.

Metric	$H(\pi)$	$V(\pi)$	Std(s)	Mean(s)
Groundtruth [22]	6.994	2.124	1.166	1.254
SJTU ours	9.608	3.745	0.428	0.769

5.4. Performance of the scanpath generation model

The metric [52] is used to evaluate the scanpath model. Table 6 gives the grand challenge related results for scanpath generation. Toolbox [56] is used and the VecSim [52] is employed as the metric. Experimental results for models including the model designed by Insight Center for Data Analytics [69] and WHU [64] are listed for comparisons. And our model gets the second place. We assume that the duration of each fixation has a positive correlation with the saliency value at its position. But experimental results show that after logistic regression, the Pearson linear correlation coefficient of the saliency value and duration is less than 0.1. The statistics show that the duration correlates with the number of viewer's fixations in single image. Besides, we calculate the rotation angles from one fixation to the next. The normalized histogram shows that most rotation angles are less than one radian, and the distribution peaks around 0.3 radian. There are some scanpaths listed in Fig. 7 in which different color indicates different scanpaths. We divide the movements along the scanpath into horizontal and vertical movements. The Table 7 gives the averaged rotation angle in the two

directions in radians, the unit of which is π. H and V are the horizontal and vertical rotation angle, respectively. The results show that the movement in horizontal direction is more frequent. We also calculate the mean and standard deviation of the duration of fixations. The results in Table 7 show that the duration of fixations have a certain degree of dispersion. Considering all the factors, we refine our model and the $VecSim$ [52] equals 4.657 for our scanpath predictor.

6. Conclusion

In this paper, we have presented a model to predict the head motions, head–eye motions and scanpaths. In our model, four different features are employed including rare spectrum in subband domain, areas of high color contrast, symmetry map and object detections. Besides, we also propose a multi-view projection and assume that there exists an equator bias when viewers are watching in HMDs. The experimental results demonstrate the effectiveness of these features, and their combination which form the overall model. In our future work, we will incorporate more visual features including both bottom-up and top-down features to improve the performance of our models. Besides, some related tasks including quality assessment and projection methods will also be explored. An effective projection method that mimics the perception in HMDs can help predict the saliency. As for quality prediction, the predicted saliency depicts the relative importance of different areas, and it can be employed to improve the quality prediction.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant 61422112, Grant 61371146, Grant 61521062, and Grant 61527804.

References

[1] Webpage, After mixed year mobile ar to drive $108 billion vr/ar market by 2021 (2017), https://www.digi-capital.com/news/2017/01/after-mixed-year-mobile-ar-to-drive-108-billion-vrar-market-by-2021/#.WbaURsig9hF.
[2] M.L. Champel, S. Lasserre, The special challenges of offering high quality experience for vr video, in: Technical Conference and Exhibition, Smpte, 2017, pp. 1–10.
[3] L. Itti, A. Borji, Computational models: Bottom-up and top-down aspects, CoRR abs/1510.07748. URL http://arxiv.org/abs/1510.07748.
[4] X. Min, G. Zhai, K. Gu, X. Yang, Fixation prediction through multimodal analysis, ACM Trans. Multimed. Comput. Commun. Appl. 13 (1) (2017) 6:1–6:23.
[5] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, W. Lin, Unified blind quality assessment of compressed natural, graphic, and screen content images, IEEE Trans. Image Process. 26 (11) (2017) 5462–5474.
[6] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, C.W. Chen, Blind quality assessment based on pseudo reference image, IEEE Trans. Multimed (2018).
[7] X. Min, G. Zhai, K. Gu, Y. Liu, X. Yang, Blind image quality estimation via distortion aggravation, IEEE Trans. Broadcast. (2018).
[8] X. Min, K. Gu, G. Zhai, M. Hu, X. Yang, Saliency-induced reduced-reference quality index for natural scene and screen content images, Signal Process. 145 (2018) 127–136.
[9] C. Grunheit, A. Smolic, T. Wiegand, Efficient representation and interactive streaming of high-resolution panoramic views, in: Proceedings. International Conference on Image Processing, Vol. 3, 2002, pp. III–209–III–212.
[10] K.-T. Ng, S.-C. Chan, H.-Y. Shum, Data compression and transmission aspects of panoramic videos, IEEE Trans. Circuits Syst. Video Technol. 15 (1) (2005) 82–95.
[11] I. Bauermann, M. Mielke, E. Steinbach, H. 264 based coding of omnidirectional video, Comput. Vis. Graph. (2006) 209–215.
[12] C.W. Fu, L. Wan, T.T. Wong, C.S. Leung, The rhombic dodecahedron map: An efficient scheme for encoding panoramic video, IEEE Trans. Multimed. 11 (4) (2009) 634–644.
[13] H. Kimata, S. Shimizu, Y. Kunita, M. Isogai, Y. Ohtani, Panorama video coding for user-driven interactive video application, in: IEEE International Symposium on Consumer Electronics, 2009, pp. 112–114.
[14] K.T. Ng, S.C. Chan, H.Y. Shum, Data compression and transmission aspects of panoramic videos, IEEE Trans. Circuits Syst. Video Technol. 15 (1) (2005) 82–95.
[15] P. Frossard, Low bit-rate compression of omnidirectional images, in: Conference on Picture Coding Symposium, 2009, pp. 53–56.
[16] H.G. Debarba, S. Perrin, B. Herbelin, R. Boulic, Embodied interaction using non-planar projections in immersive virtual reality, in: ACM Symposium on Virtual Reality Software and Technology, 2015, pp. 125–128.
[17] J. Ardouin, A. Lecuyer, M. Marchal, E. Marchand, Navigating in Virtual Environments with 360 Omnidirectional Rendering, IEEE Computer Society, 2013, pp. 95–98.
[18] J. Ardouin, A. Lcuyer, M. Marchal, E. Marchand, Stereoscopic rendering of virtual environments with wide field-of-views up to 360, in: 2014 IEEE Virtual Reality, 2014, pp. 3–8.
[19] V. Zakharchenko, K.P. Choi, J.H. Park, Quality metric for spherical panoramic video, in: Optics and Photonics for Information Processing X, 2016, 99700C.
[20] R. Carroll, M. Agrawal, A. Agarwala, Optimizing content-preserving projections for wide-angle images, ACM Trans, Graph. 28 (3) (2009) 1–9.
[21] L. Zelnik-Manor, G. Peters, P. Perona, Squaring the circle in panoramas, in: Tenth IEEE International Conference on Computer Vision, Vol. 2, 2005, pp. 1292–1299.
[22] Y. Rai, P.L. Callet, A dataset of head and eye movements for 360° images, in: ACM on Multimedia Systems Conference, 2017, pp. 205–210.
[23] W.C. Lo, C.L. Fan, J. Lee, C.Y. Huang, K.T. Chen, C.H. Hsu, 360 video viewing dataset in head-mounted virtual reality, in: ACM on Multimedia Systems Conference, 2017, pp. 211–216.
[24] C.L. Fan, J. Lee, C.Y. Huang, C.Y. Huang, K.T. Chen, C.H. Hsu, Fixation prediction for 360 video streaming in head-mounted virtual reality, in: The Workshop on Network and Operating Systems Support for Digital Audio and Video, 2017, pp. 67–72.
[25] A.D. Abreu, C. Ozcinar, A. Smolic, Look around you: Saliency maps for omnidirectional images in vr applications, in: Ninth International Conference on Quality of Multimedia Experience, 2017, pp. 1–6.
[26] Y. Rai, P.L. Callet, P. Guillotel, Which saliency weighting for omni directional image quality assessment? in: International Conference on Quality of Multimedia Experience, 2017, pp. 1–6.
[27] H.N. Hu, Y.C. Lin, M.Y. Liu, H.T. Cheng, Y.J. Chang, M. Sun, Deep 360 pilot: Learning a deep agent for piloting through 360 sports video, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2017, p. 3.
[28] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, G. Wetzstein, How do people explore virtual environments? arXiv:arXiv:1612.04335v2.
[29] E. Upenik, M. Rerabek, T. Ebrahimi, A testbed for subjective evaluation of omnidirectional visual content, in: 2016 Picture Coding Symposium, 2016, pp. 1–5.
[30] H. Duan, G. Zhai, X. Yang, D. Li, W. Zhu, Ivqad 2017: An immersive video quality assessment database, in: 2017 International Conference on Systems, Signals and Image Processing, 2017, pp. 1–5.
[31] S. Leorin, L. Lucchese, R.G. Cutler, Quality assessment of panorama video for videoconferencing applications, in: 2005 IEEE 7th Workshop on Multimedia Signal Processing, 2005, pp. 1–4.
[32] M. Yu, H. Lakshman, B. Girod, A framework to evaluate omnidirectional video coding schemes, in: IEEE International Symposium on Mixed and Augmented Reality, 2015, pp. 31–36.
[33] R. Rosenholtz, A simple saliency model predicts a number of motion popout phenomena, Vis. Res. 39 (19) (1999) 3157–3163.
[34] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.
[35] A. Torralba, A. Oliva, M.S. Castelhano, J.M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search., Psychol. Rev. 113 (4) (2006) 766.
[36] P. Reinagel, A.M. Zador, Natural scene statistics at the centre of gaze, Netw. (Bristol, Engl.) 10 (4) (1999) 341.
[37] J.H. Reynolds, R. Desimone, Interacting roles of attention and visual salience in v4, Neuron 37 (5) (2003) 853–863.
[38] M.M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, S.M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[39] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: IEEE International Conference on Computer Vision, 2010, pp. 2106–2113.
[40] O.L. Meur, P.L. Callet, D. Barba, D. Thoreau, A coherent computational approach to model bottom-up visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 28 (5) (2006) 802–817.
[41] Pelli, G. Denis, Human perception of objects: Early visual processing of spatial form defined by luminance, color, texture, motion, and binocular disparity, Optom. Vis. Sci. 78 (11) (2001) 779.
[42] E. Peli, L.E. Arend, G.M. Young, R.B. Goldstein, Contrast sensitivity to patch stimuli: effects of spatial bandwidth and temporal presentation, Spat. Vis. 7 (1) (1993) 1–14.
[43] J.M. Wolfe, T.S. Horowitz, What attributes guide the deployment of visual attention and how do they do it?, Nat. Rev. Neurosci. 5 (6) (2004) 495–501.
[44] Y.W. Guo, M. Liu, T.T. Gu, W.P. Wang, Improving photo composition elegantly: considering image similarity during composition optimization, Comput. Graph. Forum 31 (7pt2) (2012) 2193–2202.
[45] A. Jahanian, J. Liu, Q. Lin, D. Tretter, E. O'Brien-Strain, S.C. Lee, N. Lyons, J. Allebach, Recommendation system for automatic design of magazine covers, in: International Conference on Intelligent User Interfaces, 2013, pp. 95–106.
[46] L. Liu, R. Chen, L. Wolf, D. Cohen-Or, Optimizing photo composition, Comput. Graph. Forum 29 (2) (2010) 469–478.
[47] X. Tang, W. Luo, X. Wang, Content-based photo quality assessment, IEEE Trans. Multimed. 15 (8) (2013) 1930–1943.
[48] B. Kandemir, Z. Zhou, J. Li, J.Z. Wang, Beyond saliency: Assessing visual balance with high-level cues, in: Proceedings of the on Thematic Workshops of ACM Multimedia, 2017, pp. 26–34.

[49] I.C. Mcmanus, K. Stver, D. Kim, Arnheim's gestalt theory of visual balance: Examining the compositional structure of art photographs and abstract images, I-Perception 2 (6) (2011) 615–647.
[50] S. Tsogkas, I. Kokkinos, Learning-Based Symmetry Detection in Natural Images, Springer Berlin Heidelberg, 2012, pp. 41–54.
[51] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[52] O.L. Meur, T. Baccino, Methods for comparing scanpaths and saliency maps: strengths and weaknesses, Behav. Res. Methods 45 (1) (2013) 251.
[53] S. Lloyd, Least squares quantization in pcm, IEEE Trans. Inf. Theory 28 (2) (1982) 129–137.
[54] K. Gu, G. Zhai, W. Lin, X. Yang, W. Zhang, Visual saliency detection with free energy theory, IEEE Signal Process. Lett. 22 (10) (2015) 1552–1555.
[55] G. Zhai, X. Wu, X. Yang, W. Lin, W. Zhang, A psychovisual quality metric in free-energy principle, IEEE Trans. Image Process. 21 (1) (2012) 41.
[56] J. Gutiérrez, E. David, Y. Rai, P. Le Callet, Toolbox and Dataset for the Development of Saliency and Scanpath Models for Omnidirectional / 360° Still Images, Signal Process., Image Commun. 69 (2018) 35–42.
[57] C. Carlson, How i made wine glasses from sunflowers, http://blog.wolfram.com/2011/07/28/how-i-made-wine-glasses-from-sunflowers/.
[58] R.J. Peters, A. Iyer, L. Itti, C. Koch, Components of bottom-up gaze allocation in natural images, Vis. Res. 45 (18) (2005) 2397.
[59] M.O. Le, C.P. Le, D. Barba, Predicting visual fixations on video based on low-level visual features, Vis. Res. 47 (19) (2007) 2483–2498.
[60] T. Judd, F. Durand, A. Torralba, A benchmark of computational models of saliency to predict human fixations, MIT tech report.
[61] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, 2007, 545–552.
[62] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, A deep multi-level network for saliency prediction, in: 23rd International Conference on Pattern Recognition, IEEE, 2016, p. 3488.
[63] P. Lebreton, A. Raake, Gbvs360, bms360, prosal: Extending existing saliency prediction models from 2d to omnidirectional images, Signal Process., Image Commun. 69 (2018) 69–78.
[64] J. Ling, K. Zhang, Y. Zhang, D. Yang, Z. Chen, A saliency prediction model on 360° images using color dictionary based sparse representation, Signal Process., Image Commun. 69 (2018) 60–68.
[65] M. Startsev, M. Dorr, 360-aware saliency estimation with conventional image saliency predictors, Signal Process., Image Commun. 69 (2018) 43–52.
[66] F. Battisti, S. Baldoni, M. Brizzi, M. Carli, A feature-based approach for saliency estimation of omnidirectional images, Signal Process., Image Commun. 69 (2018) 53–59.
[67] Y. Fang, X. Zhang, A novel superpixel-based saliency detection model for 360-degree images, Signal Process., Image Commun. 69 (2018) 1–7.
[68] R. Monroy, S. Lutz, T. Chalasani, A. Smolic, SalNet360:' Saliency maps for omni-directional images with cnns, Signal Process., Image Commun. 69 (2018) 26–34.
[69] A. Reina, Marc, K. McGuinness, X. Giro-i Nietro, E. O'Connor, Noel, Scanpath and saliency prediction on 360° images, Signal Process., Image Commun. 69 (2018) 8–14 .