Modelling Attention in Panoramic Video: A Reinforcement Learning Approach

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

Abstract—Panoramic video provides great immersive and interactive experience, by enabling human to control field of view (FoV) with head movement (HM). Thus, HM plays a key role in modelling human attention on panoramic video. In this paper, we establish a database collecting subjects' HM positions on panoramic video sequences, and from this database we find that the HM data are highly consistent across subjects. We therefore propose an approach to predict HM positions on panoramic video, in the form of HM maps, i.e., pixel-wise heat maps for HM positions. To this end, we consider HM scan-paths as actions of an agent, such that deep reinforcement learning (DRL) can be applied in HM scan-path prediction to yield HM positions. In our approach, the DRL model is trained via optimizing reward in predicting directions and magnitudes of HM scan-paths. Based on the trained DRL model, the HM scan-paths can be generated by running multiple workflows of DRL, which are later integrated to produce HM maps. Finally, experimental results verify the effectiveness of our approach in predicting HM maps of panoramic video.

Index Terms—Computer Society, IEEE, IEEEtran, journal, LATEX, paper, template.

1 INTRODUCTION

DURING the past years, panoramic video [1] has been increasingly popular, due to its immersive and interactive experience. To achieve the immersive and interactive experience, human can control field of view (FoV) via wearing head mounted displays (HMD), when watching panoramic video in a range of $360^\circ \times 180^\circ$. In other words, humans are able to freely move their heads alongside the directions of longitude and latitude, to make their FoVs focus on the attractive content (see Figure 1 for an example). The content outside FoV cannot be seen by humans, i.e., without any attention. Consequently, head movement (HM) plays a key role in deploying human attention on panoramic video. HM prediction thus emerges as an increasingly important problem in modelling attention in panoramic video. Given the predicted HM, visual attention within FoV can be further modelled by the state-of-the-art saliency detection methods [2]. The same as traditional 2D video, attention model can be extensively utilized in many areas of panoramic video, such as region-of-interest (ROI) compression [3], visual quality assessment [4], rendering [5], synopsis [6], and automatic cinematography [7].

Unfortunately, few work has been proposed to model human attention on panoramic video, especially predicting the positions of HM. Benefitting from the most recent success of deep reinforcement learning (DRL) [8], this paper proposes a DRL based HM prediction (DHP) approach for modelling attention on panoramic video. HM prediction can be classified into two categories: offline and online manners. The offline HM prediction refers to modelling attention on panoramic video for multiple subjects, while the online prediction means predicting the next HM position of one subject upon the ground-truth of his/her HM positions at the current and previous frames. In this paper, our DHP approach includes both on-line and off-line HM prediction, named as offline-DHP and online-DHP, respectively.

To our best knowledge, there exists no offline work to predict HM positions of multiple subjects in viewing panoramic video. The closest work is saliency detection on 2D video [2]. The earliest approach for saliency detection was proposed by *ltti et al.* [9], in which the features of color, intensity and orientation are combined to generate the saliency map of an image. Later, *ltti et al.* [10] proposed to add two features in [9], motion and flicker contrast, for video saliency detection. Recently, several advanced approaches have been proposed for video saliency prediction. These advanced works include the earth mover's distance (EMD) approach [11] and the boolean map based saliency model (BMS) [12]. Most recently, deep learning has been successfully applied in video saliency detection, such as SALICON [13] and Liu's approach [14]. Saliency detection in 2D video assumes that humans are able to view all content of each video frame. However, this assumption does not hold for panoramic video, as subjects can only see a limited range of FoV at a single sight, rather than the full panoramic range of $360^\circ \times 180^\circ$. In fact, different regions of panoramic video are accessible to subjects via changing the positions of HM [15]. In this paper, we find that different subjects are highly consistent on HM positions, i.e., the longitude and latitude of their viewing directions in a panorama are similar. Such finding is based on establishing and analyzing a new database, which consists of 58 subjects' HM data in viewing 76 panoramic video sequences. Then, we propose the offline-DHP approach to predict the consistent HM positions on panoramic video via generating the HM map for each single frame. Here, the HM maps of panoramic videos are similar to the saliency maps of 2D video. Figure 1 demonstrates an example of the ground-truth HM map for a panoramic video frame.

Specifically, our offline-DHP approach yields the HM maps of panoramic video via predicting HM scan-paths of multiple *agents*, since subjects interactively control their HM positions along with some scan-paths according to video content. First, we find from our database that the HM scan-paths of different subjects are with high consistency. Meanwhile, subjects are normally initialized to view the center of front region in the beginning frames of panoramic

- *M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: see http://www.michaelshell.org/contact.html*
- *J. Doe and J. Doe are with Anonymous University.*

Manuscript received April 19, 2005; revised August 26, 2015.

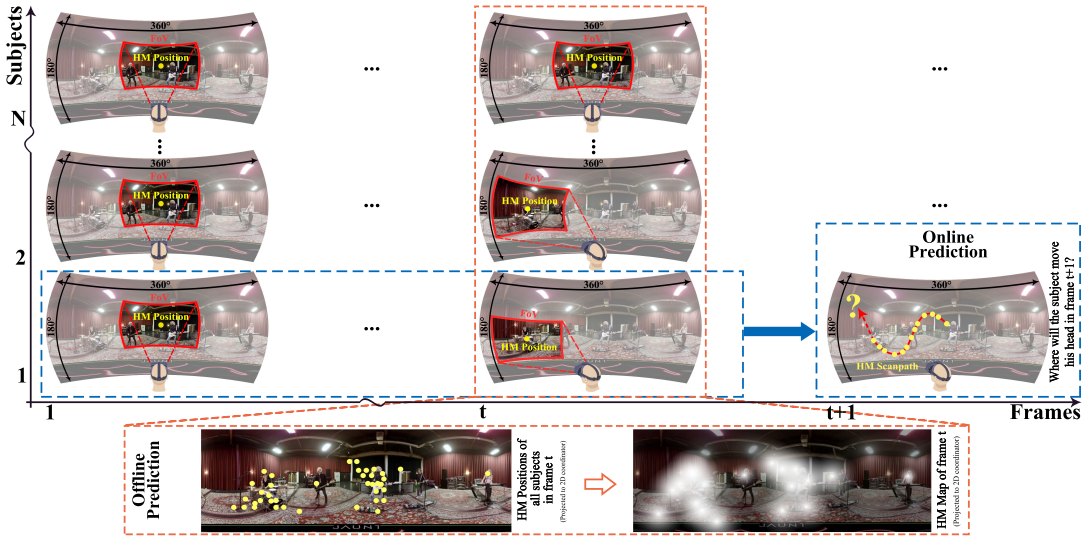


Fig. 1: Illustration for FoVs and HM positions across different subjects. The heat map of HM positions from all subjects is also shown, which is defined as the HM map.

video. Thereby, HM positions at subsequent frames can be yielded on the basis of the predicted scan-paths. Additionally, we find from our database that the magnitudes and directions of HM scan-paths are with similarity across subjects. In light of these findings, our offline-DHP approach models both magnitudes and directions of HM scan-paths as *actions* of multiple *agents* and takes viewed panoramic content as the *observation* of *environment*. As such, the DRL model of HM prediction can be learned for HM prediction. In training the DRL model, *reward* is designed to measure the difference of *actions* of HM scan-paths between the DRL *agents* and subjects. Then, the *reward* is optimized to learn parameters in the DRL model. Given the learned model, HM maps of panoramic video are generated upon HM positions, obtained from scan-paths of several *agents* in multiple DRL workflows.

For online HM prediction, the latest work of [16] proposed a deep 360 pilot, which automatically shifts viewing direction (equivalent to HM position) in watching panoramic video. Specifically, the salient object is detected and tracked across panoramic video frames, via leveraging region-based convolutional neural network (RCNN) [17] and recurrent neural network (RNN). Given the detected salient object and previous viewing directions, the deep 360 pilot predicts to transit HM position by learning a regressor. Since the deep 360 pilot relies heavily on one salient object, it is only suitable for some specific scenes that include one salient object, e.g., the sport scenes in [16]. It is still challenging to predict HM positions online for generic panoramic video, which may include more than one salient object (e.g., the panoramic video of Figure 1). In this paper, we propose an online approach, namely online-DHP, to predict HM positions on generic panoramic video. Different from [16], our online-DHP approach does not need to detect the salient object of RCNN. Instead, it is based on attention-related content by leveraging the learned model of our offline-DHP approach. Then, a DRL algorithm is developed to predict the HM positions in an online manner. Specifically, in the DRL algorithm, the *agent* predicts the *action* of HM scan-path in the next frame, according to ground-truth of the previous HM scan-path and *observation* of video content. Consequently, the HM positions at the incoming frames can be predicted for our online-DHP approach.

The main contributions of this paper are three-fold:

- We establish a new panoramic video database comprising HM positions of 58 subjects, with a thorough analysis on their HM data across 76 panoramic video sequences.
- We propose an offline-DHP approach to detect HM maps of panoramic video, which predicts the consistent HM positions of multiple subjects.
- We develop an online-DHP approach to predict the HM position of one subject at the next panoramic frame, upon video content and HM positions at the current and previous frames.

2 Related Work

2.1 Saliency detection

The only approach on predicting HM positions of panoramic videos is the most recent work of [7], in which Pano2Vid was proposed to yield FoV at each panoramic video frame. However, Pano2Vid mainly focuses on virtually generating a potential HM position at one frame, rather than modelling HM maps of multiple subjects at this frame. The closest work on predicting HM maps is saliency detection for 2D video, which is briefly reviewed in the following.

Saliency detection aims to predict visual attention of humans on 2D videos, by generating saliency maps of video frames. The studies on visual saliency start from images in 1998, when Itti and Koch [9] found that the features of intensity, color and orientation in an image can be employed to detect its saliency map. Afterwards, they extended their work to video saliency detection [10], in which two dynamic features of motion and flicker contrast are combined with [9] for detecting saliency in 2D videos. Both [9] and [10] can be seen as heuristic approaches for saliency detection, since they make use of the understanding of the HVS to develop the computational models. Recently, some advanced heuristic approaches, e.g., [11], [12], [18], [19], [20], [21], [22], [23], [24], have been proposed for saliency detection in 2D videos. Specifically, [18] proposed a novel feature called *surprise*, which

measures how the visual change attracts human observers, based on the Kullback-Leibler (KL) divergence between spatio-temporal posterior and prior beliefs. Given the feature of *surprise*, a Bayesian framework was developed in [18] for video saliency detection. Some other Bayesian frameworks [19], [20] were also developed for detecting video saliency. Besides, Lin *et al.* [11] quantified earth mover's distance (EMD) to measure the center-surround difference in spatio-temporal receptive filed, generating saliency maps for 2D videos. Zhang *et al.* [12] explored the surround cue for saliency detection, by characterizing a set of binary images with randomly thresholds on color channels. Recently, [23] and [24] have investigated that some features (e.g., motion vector) in compressed domain are of high correlation with human attention, thus being explored in video saliency detection.

Benefitting from the most recent success of deep learning, deep neural networks (DNNs) [13], [14], [25], [26], [27], [28], [29] have also been developed to detect 2D video saliency, instead of exploring the HVS related features in heuristic saliency detection approaches. They can be seen as data-driven approaches. For static saliency detection, SALICON [19] fine tuned the existing convolutional neural networks (CNN), with a new saliency related loss function. For dynamic saliency detection, [27] leveraged a deep Convolutional 3D (C3D) network to learn the representations of human attention on 16 consecutive frames, and then a Long Short-Term Memory (LSTM) network connected with a mixture density network was learned to generate saliency maps in Gaussian mixture distribution. Similarly, Liu *et al.* [14] combined CNN and multi-stream LSTM for detecting saliency in videos with multiple faces. Besides, other DNN structures have been developed to detect either static saliency [25], [26] or dynamic saliency [27], [28], [29].

Although saliency detection has been thoroughly studied for predicting eye movement on 2D videos, there is no work on the prediction of HM positions on panoramic videos. Similar to saliency detection in 2D videos, this paper proposes to generate HM maps, which represent HM positions of multiple subjects, for modelling attention on panoramic videos. Towards the HM maps of panoramic videos, a DRL approach is developed to estimate the *actions* of HM by multiple *agents* upon the *environment* of panoramic video content, the features of which are automatically learned and then extracted by DNN. Thus, our approach takes advantage of both deep learning and reinforcement learning, driven by the HM data of our panoramic video database. It is worth mentioning that although few work applies DRL to predict human attention, attention model is widely used in the opposite direction, to improve the performance of reinforcement learning, e.g., [30], [31], [32], [33].

2.2 Virtual cinematography

Virtual cinematography of panoramic videos was proposed in [7], [16], [34], [35], [36], which directs an imaginary camera to virtually capture natural FOV (NFOV). In general, virtual cinematography attempts to agree with HM positions of humans at each panoramic video frame. The early work of [34] proposed cropping object-of-interest in panoramic videos, such that NFOV can be generated for virtual cinematography. Later, in [35] the cropped object-of-interest is tracked across frames by a Kalman filter, for automatically controlling virtual camera in virtual cinematography of panoramic videos. The approach of [35] can work on both compressed and uncompressed domains, as two methods were developed for detecting object-of-interest in compressed and uncompressed domains, respectively. Both the works of [34], [35] were designed for the task of online virtual cinematography. They can be seen as heuristic approaches, which are not trained or even evaluated on the ground-truth HM data of humans.

Most recently, data-driven approaches boost the development of virtual cinematography for panoramic videos. Specifically, Pano2Vid [7] learns to generate NFOV at each panoramic frame. However, the learning mechanism of Pano2Vid is offline. In fact, NFOV can be estimated at each frame in an online manner, which uses observed HM positions of the previous frames to correct the estimation of NFOV at the current frame. To this end, online virtual cinematography has been studied [16], [36] in a data-driven way, in which HM positions are predicted by an online mechanism. Specifically, the deep 360 pilot was proposed in [16], which is a deep learning based *agent* smoothly tracking object-of-interest for panoramic video. In other words, the *agent* transits the HM position across video frames to track the key object detected by RCNN, given the observed HM position at previous frames. Consequently, the NFOV can be generated online for automatically displaying object-of-interest in virtual cinematography of panoramic videos. In fact, object-of-interest tracking in panoramic videos refers to continuously focusing and refocusing intended targets, respectively. Both focusing and re-focusing require a subject to catch up the object. Such a task is challenging in extreme-sport videos, as the object-of-interest may be moving fast. Therefore, Lin *et. al* [36] investigated two focus assistance techniques to help the subject track the key object in viewing panoramic videos, in which the potential HM position attended to the object-of-interest needs to determined and provided for the subject.

The above approaches of [7], [16], [34], [35], [36] all depend on the detector of object-of-interest. Thus, they can be only applied for some specific panoramic videos with salient objects, such as video conferencing or classroom scenes in [34], [35] and the sports videos in [7], [16], [36]. Different from these conventional approaches, our online HM prediction approach is based on the learned model of our offline approach, which encodes HM related content rather than detecting object-of-interest. Consequently, our approach is object-free, thus more suitable for the generic panoramic videos.

3 Database Establishment and Analysis

3.1 Database establishment

In this section, we collect a new database including 76 panoramic video sequences with HM data of 58 subjects, called PVS-HM database. Our PVS-HM database allows quantitative analysis of human's HM on panoramic video, and it can be also used for learning to predict where human looks at panoramic video. Our database is available in (XXX website XXX) for facilitating the future research. In the following, we present how we conducted the experiment to obtain the PVS-HM database.

First, we selected 76 panoramic video sequences from YouTube and VRCun, with resolution ranging from 3K to 8K. As seen in Table 1, the content of these sequences are diverse, including computer animation (CA), driving, action sports, movie, video game, scenery, etc. Then, the duration of each sequence was cut to be 10 to 80 seconds (averagely XXX seconds), such that fatigue can be reduced in viewing panoramic video. To ensure video quality, all panoramic video sequences were compressed by H.265 [37] without any change at bit-rates. Note that the audio tracks were removed to avoid the impact of acoustic information on visual attention.

In our experiment, 58 subjects (41 males and 17 females, aging from 18 to 36) weared the HMD of HTC Vive to view all 76 panoramic video sequences at random display order. When viewing panoramic video, the subjects seated on a swivel chair were allowed to turn around freely, such that all panoramic regions are accessible. To avoid eye fatigue and motion sickness, the subjects have a 5 minute rest after viewing each session of 19 sequences. With the support of the software development kit (SDK) of HTC Vive, we recorded the posture data of each subject when viewing panoramic video. Based on the recorded posture data, HM data of all 58 subjects at each frame of the panoramic video sequences were obtained and stored for our PVS-HM database, in terms of longitude and latitude of viewing directions.

3.2 Database analysis

In this section, we mine our PVS-HM database to analyze HM data of different subjects across panoramic video sequences. Specifically, we have the following five findings.

Finding 1: When watching panoramic video, different subjects are highly consistent in HM positions.

Analysis: In our PVS-HM database, we randomly divide all 58 subjects into two equal-size groups, A and B. For each frame of 76 sequences, the ground-truth HM maps of Groups A and B are generated by convolving with a 2D Gaussian filter along with longitude and latitude over the collected HM data. They are denoted as H_A and H_B, respectively. Note that the HM maps of H_A and H_B are in geographic coordinates. [1] [38]. For a panoramic frame, we quantify the correlation of HM maps between H_A and H_B using linear correlation coefficient (CC) [39]. Table 1 lists the averaged CC ($\pm$ standard deviations) of HM maps between Groups A and B, over all frames for each sequence. It can be seen from this table that the CC values are rather high across different sequences. It can be also seen from this table that the CC value averaged over all 48 panoramic sequences is 0.745, with the standard deviation being 0.114. Besides, Gaussian and uniform XXX. Thus, it is obvious that HM positions of subjects are highly consistent. This completes the analysis of *Finding 1*.

Finding 2: The magnitude of HM scan-paths is similar across subjects for the same panoramic content.

Analysis: HM scan-paths of human in viewing panoramic video can be decomposed into magnitude and direction. Here, we measure the magnitude of HM scan-paths in our PVS-HM database. For each individual sequence, Figure 2 shows mean and standard deviation of HM scan-path magnitude over all 58 subjects. From this figure, we can see that the standard deviation of 22.77655172 degree per second is far less than the mean of 47.71891872 degree per second, for scan-path magnitudes of HM from different subjects. Thus, there exists similarity for magnitudes of HM scan-paths across subjects, given the same panoramic video. Finally, *Finding 2* can be verified.

Finding 3: The directions of HM scan-paths on panoramic video are with high consistency across subjects.

Analysis: In each panoramic sequence, we measure the consistency of HM scan-path directions among 58 subjects of our PVS-HM database as follows. *Finding 1* has shown that the HM positions of different subjects are highly consistent when viewing panoramic video. We thus evaluate the scan-path directions of subjects starting from the consistent HM regions (i.e., regions with similar HM positions). In our PVS-HM database, such consistent regions are extracted from 76 panoramic video sequences, which has HM positions of at least 8 subjects in a small *great-circle distance* range of 3° [40]. For each panoramic video sequence, Figure 3 shows the circular standard deviation [41] of HM scan-path directions, starting from consistent HM regions within one-second time slot [2]. We can see from this figure that the circular standard deviation of HM scan-path directions is averagely 38.10° for all 76 sequences, and it is far less than 103.9° of randomly generated HM scan-path directions. This implies that there exists high consistency on directions of HM scan-paths across subjects, when viewing panoramic video. Finally, *Finding 3* can be validated.

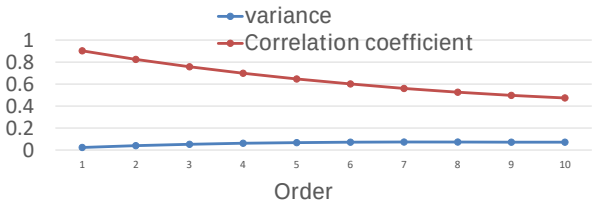


Fig. 5: Correlation coefficient of HM scan-paths over different time intervals and its variance.

Finding 4: Almost 50% subjects are consistent with one HM scan-path direction (among 8 discrete directions), and over 90% subjects are consistent with three directions for HM scan-paths.

Analysis: We measure distribution of HM scan-path directions as follows. Only HM scan-paths falling into consistent HM regions (as mentioned in *Finding 3*) are used in the following analysis. Specifically, We discretize continuous $0-360^\circ$ directions of HM scan-paths by 8-level uniform quantization: $\{0^\circ, 45^\circ, 90^\circ \cdots, 315^\circ\}$. Then, we count the proportions of subjects, whose HM scan-paths belong to the same discretized direction. Next, we rank such proportions by their values in each extracted HM region. For each ranking, the corresponding proportions of subjects are averaged over all HM regions of a panoramic sequence. Figure 4 shows the ranked proportions of subjects for each panoramic sequence. As seen in this figure, the directions of HM scan-paths from 57.1% subjects belong to the first ranked direction. In addition, the directions of HM scan-paths from 20.6% and 9.6% subjects fall into the second and third ranked directions. This completes the verification of *Finding 4*.

Finding 5: HM scan-path is predictable.The behavior of the next moment is of great relevance to the state of the last moment.We have proved that correlation between the direction of the next moment and the direction of this moment is more than 87.9% (The longer the time interval, the smaller the correlation will be).

Analysis: We used the time interval as an independent variable,calculated the correlation between t and t-1,t-2... until t-10.The results are as follows. From 5 we can see that the first-order correlation of the HM scan-paths direction is 87.9%.This indicates that the HM scan-paths direction depends largely on the direction of the last moment.That is , each HM scan-paths direction is based on the HM scan-paths direction of the last moment.And the correlation coefficient decreases with the increase of order, indicating that the correlation will decrease with the time interval increase. As seen in this figure, First order correlation coefficient up to 87.9%,while the

1. The dimension elevation is unnecessary. Here, x stands for longitude and y stands for latitude.

2. We have conducted our experiment with time slot of 0.1, 1 and 2 seconds, and the results are similar.

TABLE 1: CC between ground-truth HM maps of Groups A and B, for each panoramic video sequence

Category	Name	CC	Category	Name	CC	Category	Name	CC	Category	Name	CC
CA	AcerPredator	0.839±0.087	Driving	AirShow	0.783±0.078	Others	A380	0.839±0.106	Video Game	CS	0.819±0.084
	BFG	0.644±0.146		DrivingInAlps	0.857±0.071		CandyCarnival	0.723±0.094		Dota2	0.714±0.103
	CMLauncher	0.828±0.119		F5Fighter	0.592±0.126		MercedesBenz	0.592±0.133		GalaxyOnFire	0.762±0.084
	Cryogenian	0.526±0.174		HondaF1	0.872±0.053		RingMan	0.897±0.054		LOL	0.724±0.097
	LoopUniverse	0.779±0.078		Rally	0.867±0.047		RioOlympics	0.624±0.123		MC	0.726±0.115
	Pokemon	0.607±0.182		Supercar	0.854±0.064		VRBasketball	0.770±0.105		SuperMario64	0.860±0.054
Movie	Help	0.859±0.122	Scenery	Antarctic	0.674±0.135	Show	BTSRun	0.867±0.061	Action Sports	Gliding	0.528±0.158
	IRobot	0.771±0.078		BlueWorld	0.559±0.156		Graffiti	0.807±0.100		Parachuting	0.628±0.157
	Predator	0.696±0.124		Dubai	0.646±0.133		KasabianLive	0.722±0.132		RollerCoaster	0.834±0.078
	ProjectSoul	0.918±0.053		Egypt	0.665±0.131		NotBeAloneTonight	0.587±0.131		Skiing	0.766±0.104
	StarWars	0.950±0.016		StarryPolar	0.495±0.152		Symphony	0.779±0.096		Surfing	0.830±0.096
	Terminator	0.843±0.078		WesternSichuan	0.667±0.138		VRBasketball	0.770±0.105		Waterskiing	0.781±0.128
Overall		0.745± 0.114									

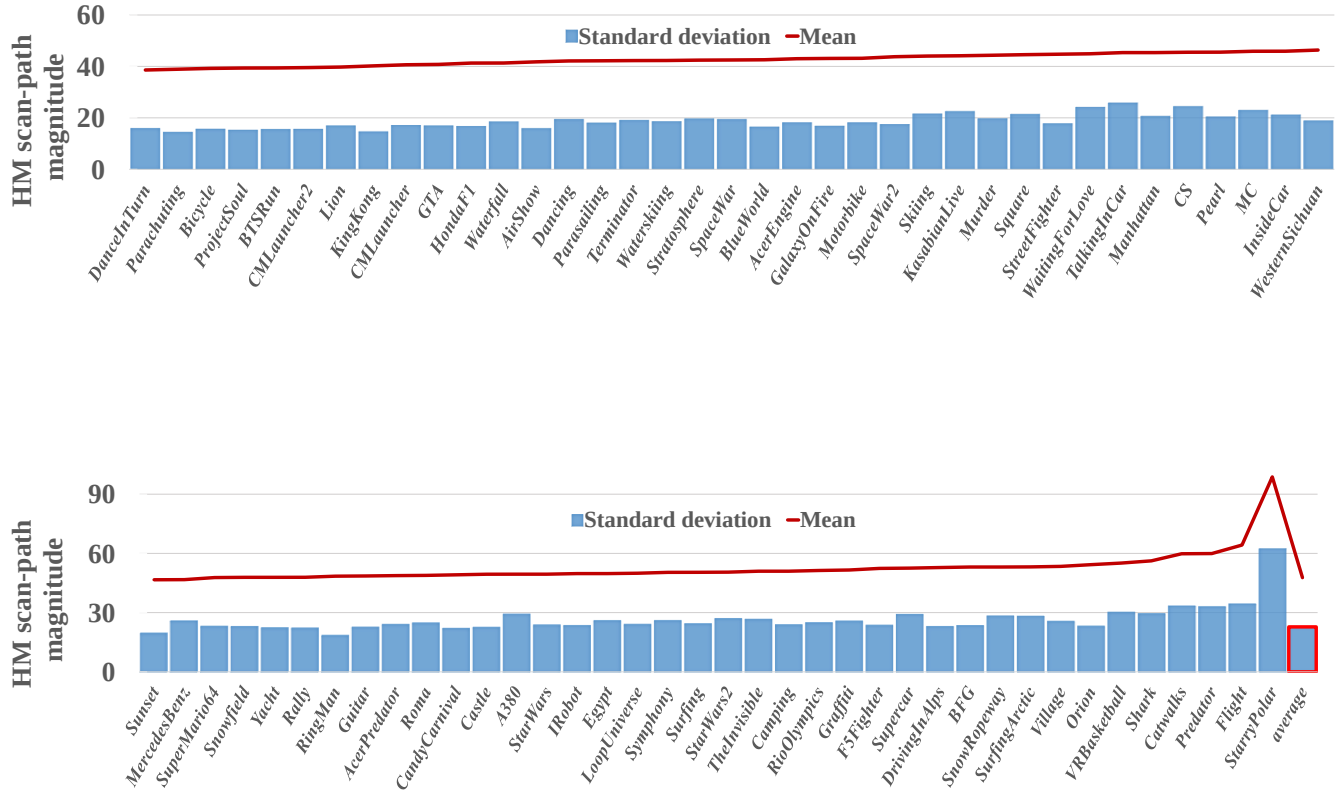
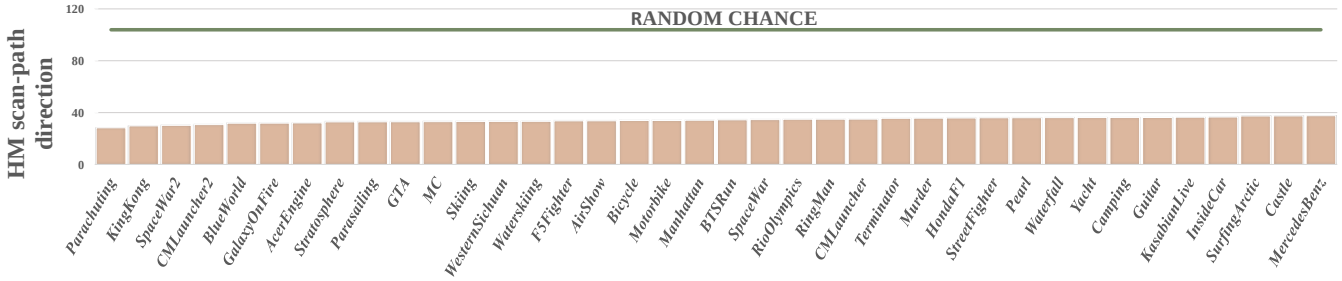


Fig. 2: Mean and standard deviation for HM scan-path magnitude across subjects, for each panoramic video sequence.



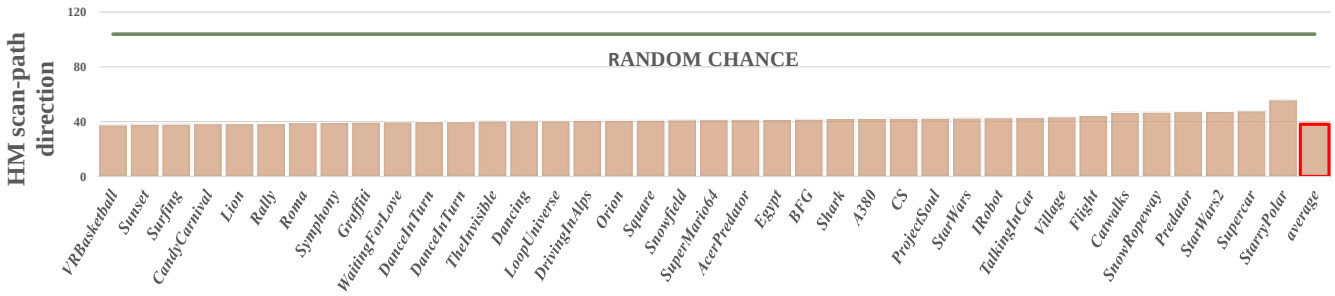


Fig. 3: Circular standard deviation for directions of HM scan-paths, over each panoramic video sequence in the HMD-VP database.

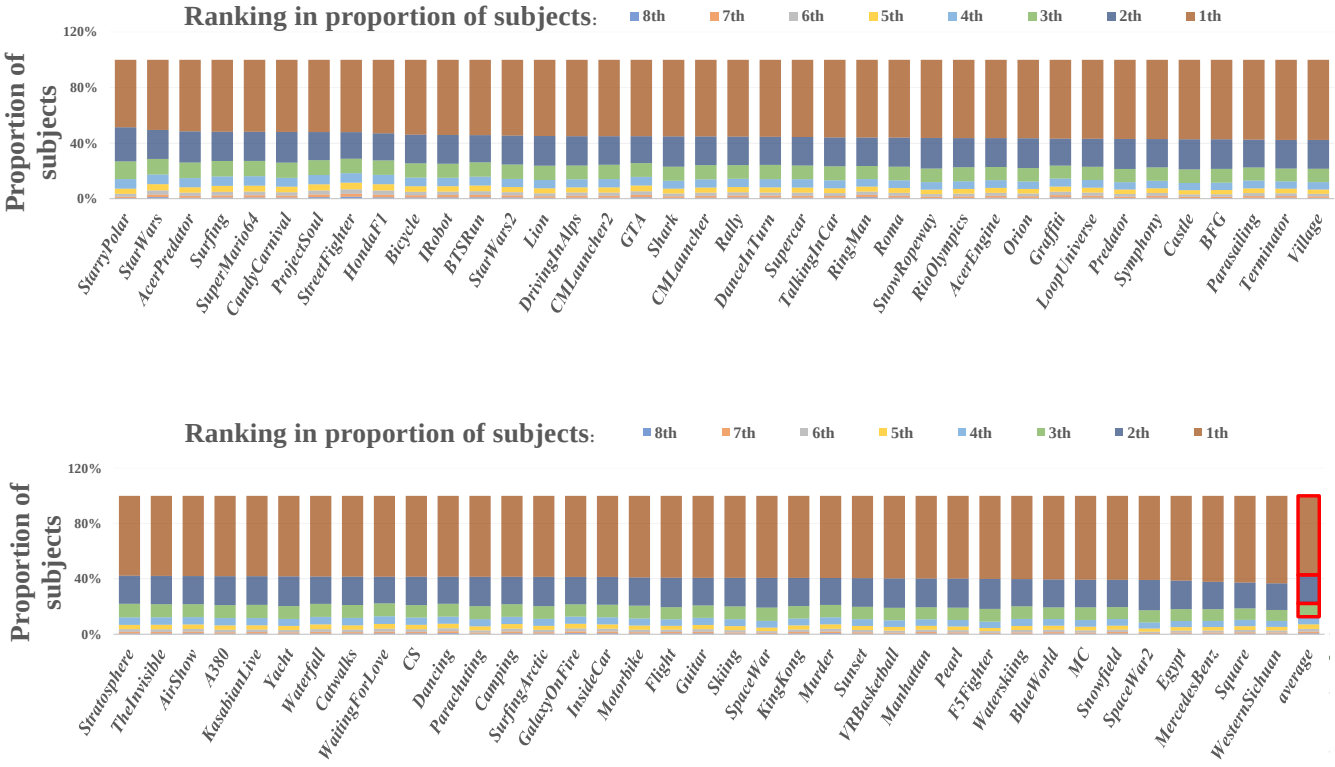


Fig. 4: Distribution of HM scan-paths over 8 discrete directions. Note that the proportions of subjects, whose HM scan-path directions fall into each discrete direction, are ranked and then shown in this figure.

variance is only 0.023. This completes the verification of *Finding 5*.

4 Offline-DHP Approach

4.1 Framework of offline-DHP

In the following, we present our offline-DHP approach, in light of our findings in Section 3.2. Figure 6 shows the overall framework of our approach, in which multiple DRL workflows are embedded to generate HM maps of the input panoramic video frames. The notations used in this figure are listed in Table 2.

As shown in Figure 6, the input to our approach is panoramic video frames $\{..., \mathbf{F}_{t-2}, \mathbf{F}_{t-1}, \mathbf{F}_t\}$. Since *Finding 1* points out that (x_t^n, y_t^n) are highly consistent for different n, it also indicates $p^n(x_t, y_t)$ consistences for different n at a specific (x_t, y_t), where . As a result, we propose to generate HM maps for modeling offline HM attention on panoramic video, which is the expected $p^n(x_t, y_t)$ across different n,

$$p(x_t, y_t) = \mathbb{E}\{p^n(x_t, y_t)\} \tag{1}$$

Above (1) is seen as the output of our approach, and is visualized with heatmap.

According to Proposition 1, $p^n(x_t, y_t)$ in (1) can be modeled by predicting a series of $\alpha_1, \alpha_2, ..., \alpha_t$ and $\nu_1, \nu_2, ..., \nu_t$. Furthermore, *Findings 2* and *3* point out predicting them is reasonable since their consistence.

Proposition 1. Assume that Ψ_{x_t,y_t} denotes all possible HM scan-path arriving at (x_t, y_t). Given the direction and magnitude of HM scan-path at frame t, i.e., α_t and ν_t, $p^n(x_t, y_t)$ can be

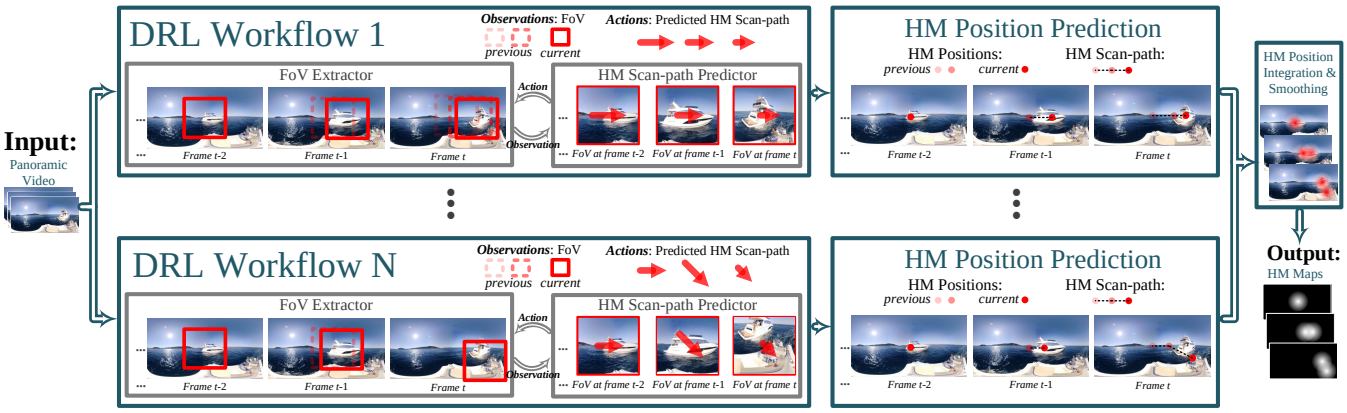


Fig. 6: Framework of our Offline-DHP approach.

TABLE 2: Notations denoted in Figure 6

$\{t\}_1^T$	The frames of panoramic video from 1 to T, with t being the currently processed frame
$\{\mathbf{F}_t\}_1^T$	The panoramic frames from 1 to T, as the input to offline-DHP/online-DHP
$\{\mathbf{o}_t\}_1^T$	The FoV from 1 to T, as the *observation* of DRL *agent*
$\{n\}_1^N$	The *subjects* from 1 to N, with n being the n-th subject
$\{i\}_1^I$	The DRL workflow from 1 to I, with i being the i-th workflow
(x_t, y_t)	The predicted HM position at frame t
(x_t^n, y_t^n)	The ground-truth HM position of n-th subject at frame t
π_t	The predicted probability distribution of HM direction at frame t, as the *policy* of DRL *agent*
α_t	The predicted HM direction at frame t, as the *action* of DRL *agent*
α_t^n	The ground-truth HM direction of the n-th subject at frame t
ν_t	The predicted HM magnitude at frame t, as the *action* of DRL *agent*
ν_t^n	The ground-truth HM magnitude of n-th subject at frame t
r_t^α	Estimated reward regarding to the predicted HM direction α_t, as the *reward* of DRL *agent*
r_t^ν	Estimated reward regarding to the predicted HM direction ν_t, as the *reward* of DRL *agent*
$p^n(*)$	The probability of event $*$ happening for the n-th subjects[3].
$\mathbf{f}_{t-1}$	LSTM feature produced by DRL model at frame $t-1$. $\mathbf{F}_t$ and f_{t-1} compose a *full observed state* of DRL *agent*

written by,

$$\begin{aligned} & p^n(x_t, y_t) \\ = & \sum_{\Psi_{x_{t-1},y_{t-1}}} \sum_{\Psi_{x_{t-2},y_{t-2}}} \cdots \sum_{\Psi_{x_1,y_1}} \\ & p^n(\alpha_{t-1}, \nu_{t-1} | \Psi_{x_{t-1},y_{t-1}}) \\ & p^n(\alpha_{t-2}, \nu_{t-2} | \Psi_{x_{t-2},y_{t-2}}) \\ & \cdots p^n(\Psi_{x_1,y_1}) \end{aligned} \tag{2}$$

Proof: For one subject, since (x_t, y_t) equals to experiencing a HM scan-path that leads to (x_t, y_t),

$$p^n(x_t, y_t) = p^n(\Psi_{x_t,y_t}). \tag{3}$$

Since Ψ_{x_t,y_t} is a conditional transition under $\Psi_{x_{t-1},y_{t-1}}$, we have

$$\Psi_{x_t,y_t} = \bigcup_{x_{t-1},y_{t-1}} \{(\alpha_{t-1}, \nu_{t-1}), \Psi_{x_{t-1},y_{t-1}}\}. \tag{4}$$

According to law of total probability, the following holds,

$$\begin{aligned} & p^n(x_t, y_t) \\ = & p^n(\Psi_{x_t,y_t}) \\ = & \sum_{\Psi_{x_{t-1},y_{t-1}}} p^n(\alpha_{t-1}, \nu_{t-1} | \Psi_{x_{t-1},y_{t-1}}) p^n(\Psi_{x_{t-1},y_{t-1}}) \end{aligned} \tag{5}$$

Above (5) can be iterated till $p^n(\Psi_{x_1,y_1})$, which derives (2) in Proposition 1.

Then, $\mathbb{E}$ in (1) can be modeled by running N DRL workflows according to Proposition 2. As a result, out offline-DHP method runs multiple DRL workflows, each of which works independently to generate an Ψ_{x_t,y_t}, i.e., (x_t, y_t) by random sampling actions based on a learnt policy. These (x_t, y_t) are latter integrated to produce expectation and achieve (1). *Finding 4* also indicates that $\mathbb{E}$ in (1) can not be achieved by running one DRL workflow with deterministic action, since α^n of different n are consistent in more than one directions.

Proposition 2. $\mathbb{E}$ in (1) can be modeled by running multiple DRL workflows, given a learned HM scan-path predictor.

$$\begin{aligned} & \mathbb{E}\{p^n(x_t, y_t)\} \\ = & \mathbb{E}_{\pi(x_t,y_t|\Psi_{x_{t-1},y_{t-1}}),\ldots,\pi(x_2,y_2|\Psi_{x_1,y_1})}\{\mathbb{E}\{p^n(\Psi_{x_1,y_1})\}\} \end{aligned} \tag{6}$$

Proof: Since *Findings 2-4* have revealed that α_t^n and ν_t^n are generally consistent across different n, we assume that $p(\alpha_{t-1}, \nu_{t-1}|\Psi_{x_{t-1},y_{t-1}})$ of all subjects is with the same probability distribution, denoted by $\pi(x_t, y_t|\Psi_{x_{t-1},y_{t-1}})$. That is,

$$p(\alpha_{t-1}, \nu_{t-1}|\Psi_{x_{t-1},y_{t-1}}) \sim \pi(x_t, y_t|\Psi_{x_{t-1},y_{t-1}}). \quad (7)$$

Thus, based on (5) and (7), the expectation of (**??**) can be rewritten in the following,

$$\begin{aligned} & \mathbb{E}\{p^n(x_t, y_t)\} \\ = \ & \mathbb{E}\{p^n(\Psi_{x_t,y_t})\} \\ = \ & \mathbb{E}_{\pi(x_t,y_t|\Psi_{x_{t-1},y_{t-1}})}\{\mathbb{E}\{p^n(\Psi_{x_{t-1},y_{t-1}})\}\}. \end{aligned} \quad (8)$$

Above (8) can be iterated till $p^n(\Psi_{x_1,y_1})$, which derives (6) in Proposition 2.

As can be seen in Figure 6, in a single DRL workflow one HM scan-path is generated by interaction between FoV extractor and HM scan-path predictor. Note that the extracted FoV is $103° \times 60°$, the same as the setting of HMD. Specifically, we can see from Figure 6 that FoV (red rectangles) is extracted according to the predicted *action* of HM scan-paths (red arrows) at previous video frames. Then, the content of extracted FoV works as *observation* of DRL, for predicting the next *action* of HM scan-path. The HM scan-path generated by each DRL workflow is forwarded to obtain HM positions at panoramic video frames. Afterwards, the HM positions from multiple DRL workflows are integrated, and then smoothed by a 2D Gaussian filter. Finally, HM maps of the panoramic video are obtained, which model the heat maps for probability of HM positions.

4.2 DRL model of offline-DHP

As described in Sections 4.1, the DRL workflow is a key component in our Offline-DHP framework, which targets at predicting HM scan-paths. This section presents how to train the DRL model in our Offline-DHP approach. Figure 7 shows the proposed training model of DRL, which can be used to predict HM scan-paths. As shown in Figure 7, FoV of the input panoramic video is extracted upon predicted HM scan-paths. In our approach, the predicted HM scan-paths and extracted FoV are seen as *action* and *observation* of DRL, respectively. In addition, *reward* is estimated on the basis of the ground-truth HM scan-paths, for generating an *action* at each panoramic video frame. Then, both FoV and estimated *reward* compose *environment*. In training the DRL model, *environment* interacts with the HM scan-path predictor. The interaction is achieved in our DRL model by the following procedure. (1) At frame t, the DRL model obtains current *observation* from the FoV extractor, denoted as o_t. In our work, *observation* o_t is the content of FoV ($103° \times 60°$) projected to 2D region and then down-sampled to 42×42.
(2) Current *observation* o_t and the feature of LSTM from the last frame f_{t-1} are delivered to the DRL network in the HM scan-path predictor.
(3) The DRL network produces LSTM feature f_t [42], HM scan-path magnitude $\hat{\nu}_t$ and policy $\hat{\pi}_t$. Here, $\hat{\pi}_t$ is the probability distribution over *actions* of HM scan-path directions. In our work, the DRL network contains four convolutional layers [43] and one LSTM layer [42], [44], which are used to extract spatial and temporal features, respectively. The details about the architecture of the DRL network can be found in Figure 7.
(4) Given $\hat{\pi}_t$, the HM scan-path predictor randomly selects an *action* $\hat{\alpha}_t$, with standard deviation ε to ensure exploration. Here, *action* $\hat{\alpha}_t$ models directions of the HM scan-path, including 8 discrete directions based on the ENU coordinate system [45], i.e., $\{0°, 45°, \cdots, 315°\}$.
(5) *Environment* is updated with $\hat{\nu}_t$ and $\hat{\alpha}_t$. Specifically, the FoV extractor returns a new *observation* o_{t+1} according to the current HM position, and the reward estimator returns the corresponding *reward* r_t^ν and r_t^α for estimating $\hat{\nu}_t$ and $\hat{\alpha}_t$, upon ground-truth HM scan-paths.
(6) A set of experiences $\{o_t, f_{t-1}, \hat{\nu}_t, \hat{\alpha}_t, r_t^\nu, r_t^\alpha\}$ is stored in an experience buffer for frame t.
(7) Once t meets the terminal condition T, which is the number of frames for the video, all experiences in the buffer are delivered to the optimizer for updating the DRL network.

Reward Estimation. Next, we focus on modeling the *reward*: r_t^α and r_t^ν. As has been discussed in Proposition 2, our goal is to make $\hat{\alpha}_t$ and $\hat{\nu}_t$ approach to ground-truth HM data. Thus, the *reward* r_t^α and r_t^ν can be represented by the difference from $\hat{\alpha}_t$ and $\hat{\nu}_t$ to their corresponding ground-truth. Let $\{\nu_t^n, \alpha_t^n\}$ be the ground-truth magnitude and direction of HM scan-path, and $\{x_t^n, y_t^n\}$ be the ground-truth HM position, at frame t for the n-th subject. Note that $\{x_t^n, y_t^n\}$ is the geographic coordinate position. [38]. Then, r_t^α can be written as

$$r_t^\alpha = \frac{1}{N}\sum_{n=1}^{N} e^{-\frac{1}{2}\left(\frac{D_d(\hat{\alpha}_t, \alpha_t^n)}{\rho}\right)^2} e^{-\frac{1}{2}\left(\frac{D_s((\hat{x}_t, \hat{y}_t), (x_t^n, y_t^n))}{\varrho}\right)^2}, \quad (9)$$

where D_d defines *phase difference*, D_s denotes *great-circle distance* [46] and ρ, ϱ are hyper-parameters. Similarly, we have

$$r_t^\nu = \frac{1}{N}\sum_{n=1}^{N} e^{-\frac{1}{2}(\hat{\nu}_t - \nu_t^n)^2} e^{-\frac{1}{2}\left(\frac{D_d(\hat{\alpha}_t, \alpha_t^n)}{\rho}\right)^2} e^{-\frac{1}{2}\left(\frac{D_s((\hat{x}_t, \hat{y}_t), (x_t^n, y_t^n))}{\varrho}\right)^2}. \quad (10)$$

Proposition 3. Assume that the probability of each pixel $\{\hat{x}_t, \hat{y}_t\}$ being the HM position decays along with the distance to the ground truth HM position, following 2D coordinate Gaussian distribution along with longitude and latitude. In addition, assume that the probability for the degree α_t^n of HM scan path also follows the Gaussian distribution with the mean being α_t^n. Then r_t^α can be represented by (9).

Proof: At frame t, the ground truth HM position is $\{x_t^n, y_t^n\}$ for the n-th subject. besides, α_t^n is the direction of HM scan path for the n-th subject. Thus, we can acquire:

$$P(\alpha_t^n|\{x_t^n, y_t^n\}) = 1. \quad (11)$$

Since the probability of $\{\hat{x}_t, \hat{y}_t\}$ being the HM position follows Gaussian distribution centered at $\{x_t^n, y_t^n\}$, the equality below can be obtained:

$$P_n(\alpha_t^n|\{\hat{x}_t, \hat{y}_t\}) = P_n(\alpha_t^n|\{x_t^n, y_t^n\}) e^{-\frac{1}{2}\left(\frac{D_s((\hat{x}_t, \hat{y}_t), (x_t^n, y_t^n))}{\varrho}\right)^2}. \quad (12)$$

Similarly, we have:

$$P_n(\hat{\alpha}_t|\{\hat{x}_t, \hat{y}_t\}) = P_n(\alpha_t^n|\{\hat{x}_t, \hat{y}_t\}) e^{-\frac{1}{2}\left(\frac{D_d(\hat{\alpha}_t, \alpha_t^n)}{\rho}\right)^2}. \quad (13)$$

Based on (12), (13) can be written as:

$$\begin{aligned} & P_n(\hat{\alpha}_t|\{\hat{x}_t, \hat{y}_t\}) = P_n(\alpha_t^n|\{x_t^n, y_t^n\}) \cdot \\ & e^{-\frac{1}{2}\left(\frac{D_s((\hat{x}_t, \hat{y}_t), (x_t^n, y_t^n))}{\varrho}\right)^2} e^{-\frac{1}{2}\left(\frac{D_d(\hat{\alpha}_t, \alpha_t^n)}{\rho}\right)^2}. \end{aligned} \quad (14)$$

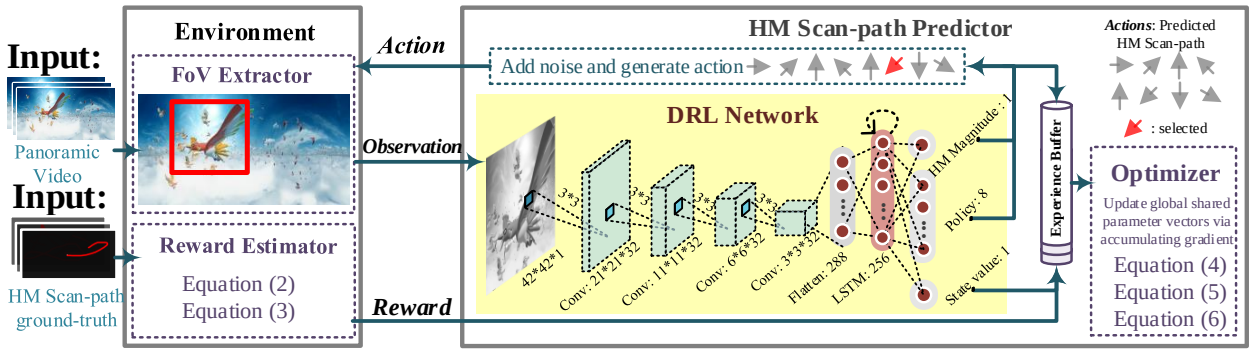


Fig. 7: DRL model for the HM scan-path predictor.

According to (11), the following holds:

$$P_n(\hat{\alpha}_t|\{\hat{x}_t,\hat{y}_t\}) = e^{-\frac{1}{2}\left(\frac{D_d(\hat{\alpha}_t,\alpha_t^n)}{\rho}\right)^2} e^{-\frac{1}{2}\left(\frac{D_s((\hat{x}_t,\hat{y}_t),(x_t^n,y_t^n))}{\varrho}\right)^2}. \tag{15}$$

Reward r_t^α can be represented by $P_n(\hat{\alpha}_t|\{\hat{x}_t,\hat{y}_t\})$ averaged overall subjects, i.e., (9),which estimates how likely human agent conducts the action of $\hat{\alpha}_t$. Finally, this proposition is proved.

Proposition 4. Assume that the probability of each pixel $\{\hat{x}_t,\hat{y}_t\}$ being the HM position decays along with the distance to the ground truth HM position, following 2D coordinate Gaussian distribution along with longitude and latitude. In addition, assume that the probability for the degree α_t^n of HM scan path also follows the Gaussian distribution with the mean being α_t^n. Then r_t^α can be represented by (9). Provided that the probability for the magnitude $\hat{\nu}_t$ of HM scan path follows the Gaussian distribution with the mean of ν_t^n. We have (10).

Proof: At frame t, ν_t^n is the magnitude of HM scan path at the direction α_t^n for the n-th subject. Then, similar to the proof of (11), the following holds:

$$P(\nu_t^n|\{x_t^n,y_t^n\},\alpha_t^n) = 1. \tag{16}$$

$$P_n(\nu_t^n|\{\hat{x}_t,\hat{y}_t,\hat{\alpha}_t\}) = P(\nu_t^n|\{x_t^n,y_t^n\},\alpha_t^n)\cdot e^{-\frac{1}{2}\left(\frac{D_s((\hat{x}_t,\hat{y}_t),(x_t^n,y_t^n))}{\varrho}\right)^2} e^{-\frac{1}{2}\left(\frac{D_d(\hat{\alpha}_t,\alpha_t^n)}{\rho}\right)^2}. \tag{17}$$

Due to the Gaussian distribution of the probability for magnitude $\hat{\nu}_t$, we can obtain:

$$\begin{aligned} & P_n(\nu_t^n|\{\hat{x}_t,\hat{y}_t,\hat{\alpha}_t\}) \\ = \; & P(\nu_t^n|\{x_t^n,y_t^n\},\alpha_t^n) e^{-\frac{1}{2}(\hat{\nu}_t-\nu_t^n)^2}\cdot \\ & e^{-\frac{1}{2}\left(\frac{D_s((\hat{x}_t,\hat{y}_t),(x_t^n,y_t^n))}{\varrho}\right)^2} e^{-\frac{1}{2}\left(\frac{D_d(\hat{\alpha}_t,\alpha_t^n)}{\rho}\right)^2}. \end{aligned} \tag{18}$$

According to (16) and (18), we have:

$$P_n(\nu_t^n|\{\hat{x}_t,\hat{y}_t,\hat{\alpha}_t\}) = e^{-\frac{1}{2}(\hat{\nu}_t-\nu_t^n)^2}\cdot e^{-\frac{1}{2}\left(\frac{D_s((\hat{x}_t,\hat{y}_t),(x_t^n,y_t^n))}{\varrho}\right)^2} e^{-\frac{1}{2}\left(\frac{D_d(\hat{\alpha}_t,\alpha_t^n)}{\rho}\right)^2}. \tag{19}$$

Reward r_t^ν can be similarly represented by $P(\nu_t^n|\{x_t^n,y_t^n\},\alpha_t^n)$ averaged overall subjects, i.e., (10). This completes the proof of proposition 4.

Optimization. Now, we need to optimize *reward*: r_t^α and r_t^ν, such that we can learn the network parameters of our DRL model in Figure 7. Our Offline-DHP approach implements the asynchronous DRL method [8]. Hence, multiple workflows are run to interact with multiple *environments* with workflow-specific parameter vectors $\{\theta_{\hat{\nu}}^{'},\theta_{\hat{\pi}}^{'},\theta_V^{'}\}$, producing $\hat{\nu}_t$, $\hat{\pi}_t$ and V. Here, V denotes *state value* output by the DRL network. Meanwhile, global-shared parameter vectors $\{\theta_{\hat{\nu}},\theta_{\hat{\pi}},\theta_V\}$[4] are updated via accumulating gradient. For more details about workflow-specific and global-shared parameter vectors, refer to [8]. In our approach, *reward* r_t^ν is optimized to train $\theta_{\hat{\nu}}$ as follows,

$$d\theta_{\hat{\nu}} \leftarrow d\theta_{\hat{\nu}} + \nabla_{\theta_{\hat{\nu}}^{'}} \sum_{t=1}^{t_{\max}} r_t^\nu. \tag{20}$$

Besides, we can optimize *reward* r_t^α by

$$d\theta_V \leftarrow d\theta_V + \nabla_{\theta_V^{'}} \sum_{t=1}^{t_{\max}} \Big(\sum_{i=t}^{t_{\max}} \gamma^{i-t} r_i^\alpha - V(o_t,f_{t-1};\theta_V^{'})\Big)^2, \tag{21}$$

$$\begin{aligned} d\theta_{\hat{\pi}} \leftarrow d\theta_{\hat{\pi}} + \nabla_{\theta_{\hat{\pi}}^{'}} \sum_{t=1}^{t_{\max}} & \log \hat{\pi}(\alpha_t|o_t,f_{t-1};\theta_{\hat{\pi}}^{'})\cdot \\ & \Big(\sum_{i=t}^{t_{\max}} \gamma^{i-t} r_i^\alpha - V(o_t,f_{t-1};\theta_V^{'})\Big), \end{aligned} \tag{22}$$

where γ is the discount factor of *Q-learning* [47]. In addition, $V(o_t,f_{t-1};\theta_V^{'})$ denotes state value V yielded by o_t, f_{t-1} and $\theta_V^{'}$; $\hat{\pi}(\alpha_t|o_t,f_{t-1};\theta_{\hat{\pi}}^{'})$ stands for the probability of *action* α_t that yields by policy $\hat{\pi}_t$ from o_t, f_{t-1} and $\theta_{\hat{\pi}}^{'}$. Finally, upon the above equations, RMSProp [48] is applied to optimize *reward* in training data. As a result, workflow-specific and global-shared parameter vectors can be learned for predicting HM scan-paths.

5 Online-DHP Approach

In this section, we present our online-DHP approach. Our online-DHP approach refers to predicting a specific subject's HM position at the next frame, given his/her HM scan-path till the current frame. Here, we define this subject as a *viewer*. Figure 8 shows the framework of our online-DHP approach, which predicts the HM

4. As can be seen in Figure 7, $\{\theta_{\hat{\nu}},\theta_{\hat{\pi}},\theta_V\}$ share all CNN and LSTM layers in our Offline-DHP approach, but they are separated at the output layer.

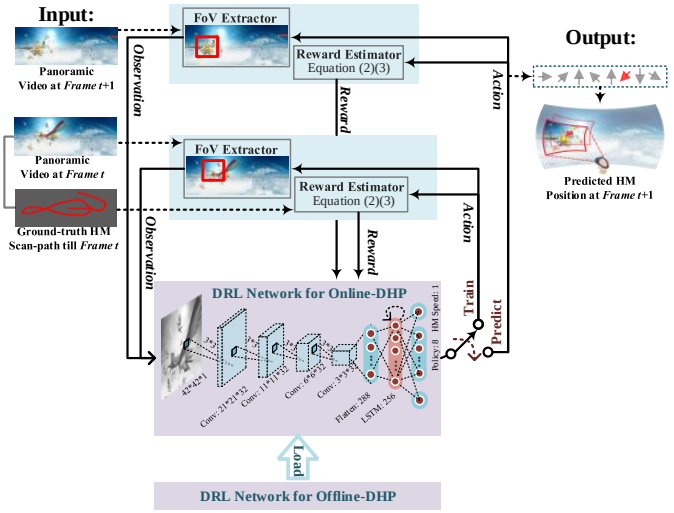


Fig. 8: Framework of our Online-DHP approach.

position of the viewer at frame $t+1$ upon observation till frame t. As shown in this framework, the input to our online-DHP approach is the viewer's HM scan-path and frame content till frame t, while the output is the predicted HM position at frame $t+1$ for the viewer. At the beginning, the HM position of the viewer is initialized to be the center of front region, which is the general setting of the panoramic video player. Then, the trained DRL network of our offline-DHP approach is loaded as the initial DRL network for online prediction, and both networks share the same structure. The reason for loading the offline-DHP network is that it contains the knowledge of HM related features. Later, this initial DRL network is fine-tuned by the viewer's HM scan-path at incoming frames.

As aforementioned, the offline DRL network is loaded in the initialization of our online-DHP approach. Thus, training the DRL network for online-DHP generally follows that of offline-DHP, both containing the basic reinforcement learning procedure. The following shows the procedure of one episode,

1) Take *action* upon current *observation*. The action selects 1 among 8 discrete HM scan-path directions, i.e., $\{0°, 45°, \cdots, 315°\}$. The action also contains a scalar of HM scan-path magnitude.
2) Calculate *reward* from the reward estimator with (9) and (10), which measures how close this action to the ground-truth action of the viewer.
3) Generate new *observation* from the FoV extractor with above *action*, and then make the content of extracted FoV as the input of the DRL network
4) Iterate from 1) to 4), if the number of iteration does not exceed t. Otherwise, update DRL network with (21) and (22). This episode ends and move to the next episode.

The detailed implementation of the above process has been presented in Section 4.2. Besides, the definition of *action*, *reward* and *observation* is the same as Section 4.2. However, the main differences between the online- and offline-DHP approach are:

- For online-DHP, the termination condition of iteration is t, where t is the frame the viewer is currently seeing, instead of the last frame of the video in offline-DHP.
- For online-DHP, the ground-truth HM scan-path used in (9) and (10) is from the single viewer, rather then from all subjects.

The above training procedure ends for frame t, once meeting the termination condition. In our approach, the termination condition is based on the mean overlap (MO), which measures how close the predicted HM position to the ground truth HM position. The MO ranges from 0 to 1, where larger MO indicates more precise prediction. Specifically, MO is defined as,

$$\mathrm{MO} = \frac{A(\mathrm{FoV}_p \cap \mathrm{FoV}_g)}{A(\mathrm{FoV}_p \cup \mathrm{FoV}_g)}, \tag{23}$$

where FoV_p and FoV_g represent the FoVs at predicted and ground-truth HM positions, respectively. In (23), A represents the area of a panoramic region, which counts for number of pixels. When the average MO is larger than a threshold 0.7 [5], the switch in Figure 8 is turned to "predict" and the DRL network make a prediction for action at frame $t+1$. Note that if the number of training episodes exceeds E, the "predict" is also switched on, so that the training episodes end in a limited time. The above prediction for action at frame $t+1$ includes the direction and magnitude of HM scan-path. As a result, the HM position at frame $t+1$ can be predicted, given the ground-truth HM position at frame t. Then, similar training procedure is conducted for frame $t+1$, for the HM position at frame $t+2$. Finally, online-DHP is achieved by training till current frame, for predicting at the next frame.

6 Experimental Results

6.1 Settings

In this section, experiments were conducted to evaluate the performance of our DHP approach. In our experiments, we randomly divided the 76 panoramic video sequences of our PVS-HM database into a training set (61 sequences) and a test set (15 sequences). For training the DRL model, we set ρ and ϱ of (9) and (10) to be 42 and 0.73 for estimating *reward* of HM scan-path prediction. Note that these two hyperparameters were tuned over the test set. In addition, we empirically choose γ of (21) and (22) to be 0.99 for *reward* optimization, which is a widely used setting in other DRL works. Besides, we followed [8] to set other hyperparameters of DRL. In this paper, all 61 training sequences, each of which corresponds to a local DRL network, were used to update the global network as the trained DRL model. For predicting HM positions, the number of DRL workflows in our approach (see the framework of Figure 6) was set to be 56, the same as the number of subjects in our PVS-HM database. For producing final HM maps of panoramic video frames, we followed the setting of [40] to convolute the predicted HM positions with a 2D Gaussian filter. In the following performance evaluation, the HM maps in a panorama were projected to the 2D coordination.

6.2 Performance evaluation

Since there is no work on predicting HM maps of panoramic video and saliency prediction is the closest field, we compare our DHP approach with three state-of-the-art saliency prediction approaches, i.e., BMS [12], SALICON [?] and OBDL [23] . In particular, OBDL [23] and BMS [12] are the latest saliency prediction approaches for video and image, respectively. We also

5. The threshold is set to the average MO achieved in [16]

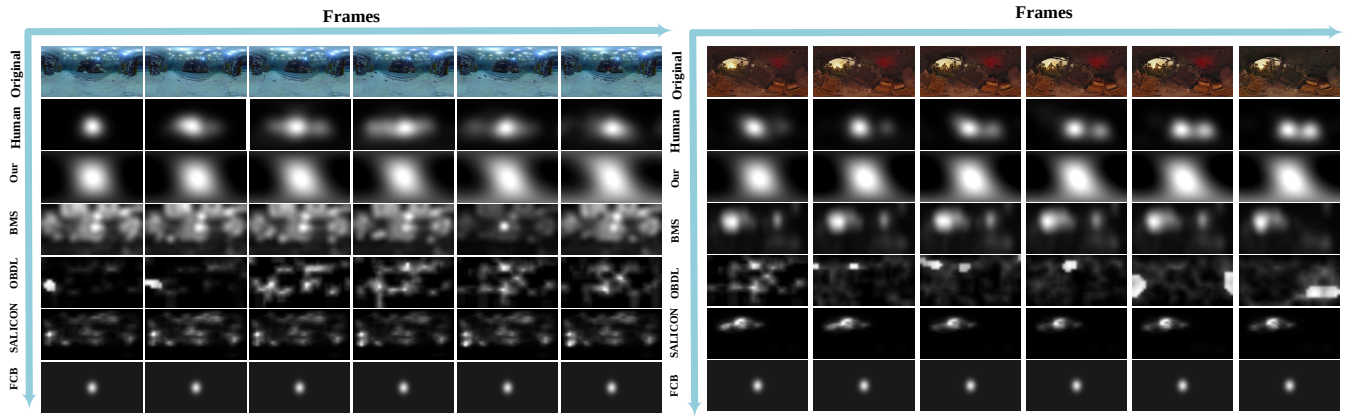


Fig. 9: HM maps of several frames selected from two test sequences in our PVS-HM database. They are all visualized in the 2D coordination. The second row shows the ground-truth HM maps, which are generated upon HM portions of all 56 subjects. The third to sixth rows show the HM maps of our, BMS [12] , OBDL [23],and SALICON approaches. The last row demonstrates the HM maps by the FCB baseline.

TABLE 3: CC results of HM map prediction by our and other approaches

CC	Method	Average	KingKong	SpaceWar2	StarryPolar	Dancing	Guitar	BTSRun	InsideCar	RioOlympics	SpaceWar	CMLauncher2	Waterfall	Sunset	BlueWorld	Symphony	WaitingForLove
Non-FCB	Our	0.808	0.711	0.815	0.807	0.821	0.771	0.775	0.811	0.834	0.751	0.837	0.863	0.874	0.782	0.841	0.832
	BMS	0.328	0.007	0.091	0.496	0.233	0.528	0.276	0.481	0.289	0.305	0.364	0.479	0.548	0.177	0.287	0.363
	OBDL	0.183	-0.010	0.099	-0.018	0.323	0.124	0.200	0.337	0.285	0.053	0.170	0.052	0.470	0.018	0.291	0.348
	SALICON	0.231	-0.044	0.059	0.355	0.115	0.425	0.196	0.550	0.097	0.169	0.243	0.152	0.206	0.308	0.326	0.304
FCB	Our	0.609	0.516	0.490	0.578	0.654	0.595	0.535	0.583	0.612	0.431	0.753	0.663	0.678	0.642	0.733	0.672
	BMS	0.578	0.463	0.450	0.495	0.601	0.587	0.430	0.568	0.607	0.392	0.746	0.653	0.646	0.651	0.713	0.675
	OBDL	0.448	0.280	0.362	0.382	0.478	0.264	0.297	0.517	0.487	0.330	0.714	0.661	0.463	0.567	0.397	0.521
	SALICON	0.466	0.142	0.260	0.454	0.377	0.478	0.398	0.538	0.461	0.347	0.640	0.523	0.595	0.614	0.572	0.592
FCB Only		0.616	0.517	0.495	0.580	0.655	0.591	0.533	0.589	0.621	0.441	0.768	0.676	0.694	0.649	0.745	0.686

compare our approach to the front-center-bias (FCB) baseline, since human attention normally biases towards front-center regions of panoramic video. Here, we model FCB by Gaussian distribution[6], similar to the center bias of saliency detection. Note that in the field of saliency detection, center bias [2] has been already verified to be effective, and it is thus multiplied to saliency maps for improving saliency detection accuracy. Hence, we further report results of HM maps multiplied by FCB, for our and other approaches. In our experiments, we evaluate accuracy of HM map prediction in terms of CC and NSS(The Normanized Scanpath Saliency) [49], which are two common evaluation metrics in saliency prediction [2].NSS measures the average salicency map values at human's fixation locations.The saliency map was normalized to have zero mean and united standard deviation, then the mean of normalized values that corresponding to the human's fixation locations was taken as the NSS score:

$$NSS = \frac{1}{N} * \sum_{p=1}^{N} \frac{S(p) - \mu_S}{\sigma_S} \tag{24}$$

Where N is the number of fixations and p is the human eye positions.NSS ≤ 0 indicates anti-correspondence between predicted saliency points and fixation loacations,while positive NSS suggest greater correspondence than chance.

6. The standard deviation is 12.77 degree, obtained by mean squared error (MSE) fitting over all HM positions of the training set.

Table 3 compares CC results of our and other approaches, in predicting HM maps over all 15 test sequences of panoramic video. Here, CC results are averaged over all frames for each test sequence. We can see from this table that our approach is far superior to other approaches, when FCB is not integrated (i.e., non-FCB). Specifically, our approach has 0.625 ,$0.577$and 0.480 CC increment, over OBDL ,SALICON,and BMS, respectively. Once integrated with FCB, all four approaches have performance improvement, and our approach still performs best among all three approaches. Additionally, our approach significantly outperforms the FCB baseline. In a word, our DHP approach is effective in predicting HM maps of panoramic video, much better than other approaches and the FCB baseline.

Table 4 shows NSS results of our and other approaches, in predicting HM maps over all 15 test sequences of panoramic video. The NSS results are also averaged over all frames for each test sequence.From this table, we can see that our approach has more outstanding performance than other approaches when the FCB is not applied.Specifically, our approach has 1.617,1.454 and 1.754 NSS increment, over OBDL ,SALICON and BMS, respectively. All four approaches have performance improvement when integrated with FCB, and our approach still scored best among all four approaches. Moreover, our approach significantly outperforms the FCB baseline. In conclusion, the NSS score Demostrate that our DHP approach is effective in predicting HM maps of panoramic

TABLE 4: NSS results of HM map prediction by our and other approaches

NSS	Method	Average	KingKong	SpaceWar2	StarryPolar	Dancing	Guitar	BTSRun	InsideCar	RioOlympics	SpaceWar	CMLauncher2	Waterfall	Sunset	BlueWorld	Symphony	WaitingForLove
Non-FCB	Our	2.074	1.516	1.452	1.452	2.542	2.087	2.399	2.426	2.021	1.507	2.676	2.344	2.047	1.986	2.246	2.406
	BMS	0.950	-0.019	0.137	1.313	0.670	1.497	0.546	1.657	0.772	0.807	1.034	1.673	1.613	0.841	0.710	0.997
	OBDL	0.457	-0.107	0.301	-0.006	0.980	0.393	0.215	1.375	0.637	0.064	0.660	0.073	1.015	0.035	0.260	0.964
	SALICON	0.620	-0.203	0.051	0.730	-0.236	0.965	0.337	1.823	0.600	0.344	0.921	0.410	0.669	1.138	0.456	1.298
FCB	Our	2.998	1.986	1.358	2.628	3.799	3.080	2.422	2.355	2.305	1.576	5.805	3.861	2.884	3.081	4.443	3.382
	BMS	2.802	1.529	1.109	2.621	3.068	3.079	1.573	2.267	2.308	1.472	5.465	4.014	2.815	3.127	4.229	3.356
	OBDL	1.965	0.580	0.916	1.618	1.947	0.834	0.564	2.138	1.947	1.100	5.080	4.060	1.814	2.572	1.465	2.844
	SALICON	2.169	0.019	0.448	2.411	0.641	2.117	1.728	2.013	1.776	1.380	5.333	3.009	2.674	2.909	2.780	3.292
FCB Only		2.130	0.927	0.634	2.300	2.023	1.744	1.063	1.172	1.416	1.090	5.736	3.122	2.259	2.125	3.994	2.348

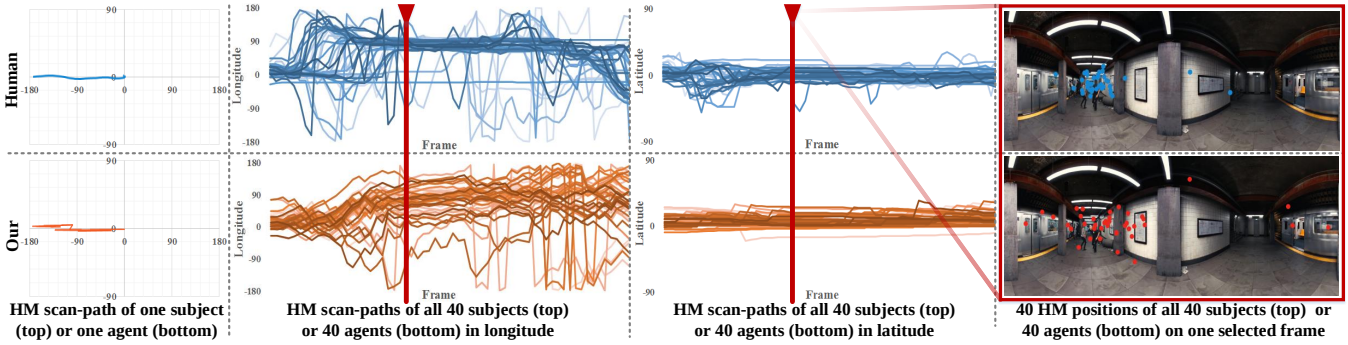


Fig. 10: Visualization in scan-paths and positions of HM generated by ground-truth and our approach, for sequence *Help*. The first column shows scan-path of one subject and one DRL workflow. The second and third columns show scan-paths of all 40 subjects and 40 DRL workflows, in longitude and latitude directions. The last column visualizes HM positions of ground-truth and our approach.

video, much better than other approaches and the FCB baseline.

Next, we move to the comparison of subjective results. In Figure 9, we visualize the HM maps of several frames from 2 selected sequences, which are generated by our, BMS and OBDL approaches as well as ground-truth HM positions and the FCB baseline. Here, the predicted HM maps are produced by three approaches with integrated FCB, as FCB can enhance performance of all three approaches (as shown in Table 3). From this figure, one can observe that HM maps of our approach are more close to ground-truth maps, compared to other approaches and the FCB baseline. This indicates that our DHP approach is capable of better locating HM positions on panoramic video. Moreover, we plot in Figure 10 the HM scan-paths by subjects and by our DHP approach, to investigate how well the proposed DRL workflows agree with the ground-truth HM scan-paths. We can see from this figure that our approach is able to yield similar scan-paths as humans. We further show in Figure 10 the HM positions obtained from those scan-paths, which are used to produce HM maps. Again, this figure reveals that HM positions can be well predicted by our approach. In a word, our DHP approach is effective in modeling both scan-paths and positions of HM.

6.3 Online-DHP performance evaluation

Now, we move to the evaluation of Online-DHP. We compare our Online-DHP approach with Deep 360 Pilot approach [16] and two baselines.

Table 5 compares mean MO results of our and other approaches, in predicting HM positions over all 15 test sequences of panoramic video. Here, MO results are averaged over all frames for each test sequence. The Random-baseline means predicting the HM position in a totally random way. And the Keep-baseline means pedicting the HM position due to the HM position of the former frame. We can see from this table that our approach performs far better than other approaches. Specifically, our approach has ? average increment over Deep 360 Pilot approach [16]. We think that is mostly because that our approach has priori knowledge from the Offline-DHP. For the two baselines, the Keep-baseline is a little better than the Random-baseline mostly because of the guidance of the former frame. Compared with the two baselines, our approach is also obviously much more effective.

Offline-DHP is an important component of our approach. Figure 11 shows the result of mean MO between Online-DHP and Online-DHP without Offline-DHP. We can see that the Offline-DHP enhances the performance of online prediction a lot. We think it's mostly because the Offline-DHP provides priori knowledge for the online prediction. And this is very effective in predicting.

After the comparison, we move to the analysis of an important parameter th_{MO} in algorithm 1, which affects both the result of MO(23) and the time complexity of our approach. We call it the threshold of MO. According to our approach, we think the threshold th_{MO} is positively correlated with the result of MO, but negatively correlated with the time complexity.

7 Discussion

Different from traditional 2D video, panoramic video offers $360° \times 180°$ scenes for immersive experience. To access preferable

TABLE 5: MO results of HM position prediction by our and other approaches

Method	KingKong	SpaceWar2	StarryPolar	Dancing	Guitar	BTSRun	InsideCar	RioOlympics	SpaceWar	CMLauncher2	Waterfall	Sunset	BlueWorld	Symphony	WaitingForLove	Average
Online	0.809	0.763	0.549	0.859	0.785	0.878	0.847	0.820	0.626	0.763	0.667	0.659	0.693	0.747	0.836	0.753
Deep 360 Polit	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	#DIV/0!
Random-baseline	0.201	0.206	0.161	0.216	0.203	0.206	0.216	0.203	0.209	0.205	0.203	0.204	0.206	0.202	0.211	0.204
Keep-baseline	0.224	0.231	0.197	0.217	0.227	0.237	0.234	0.225	0.216	0.251	0.251	0.209	0.229	0.216	0.225	0.226

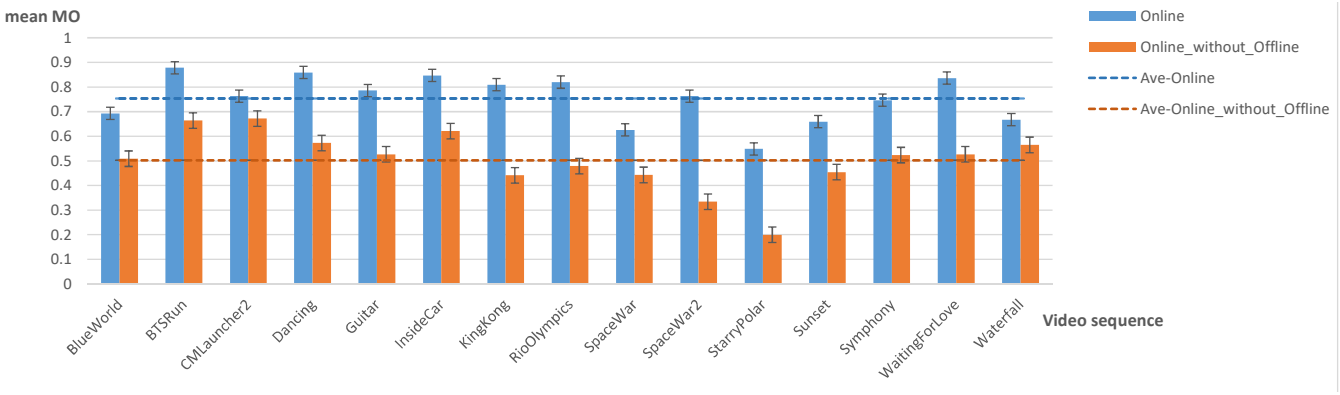


Fig. 11: MO results between Online-DHP and Online-DHP without offline-DHP.

FoV, humans need to interact with panoramic video by moving their heads. Therefore, human attention is deployed according to HM. Interestingly, we found that HM data of viewing panoramic video are similar across humans. This finding was achieved by establishing a new database called PVS-HM, which includes HM data of 40 subjects on viewing 48 panoramic video sequences. In this paper, we proposed a DHP approach to predict maps of HM positions on panoramic video. Specifically, our approach leverages DRL to predict *actions* of HM scan-paths. Afterwards, HM scan-paths of several *agents* from multiple DRL workflows are integrated to yield HM maps, which encode possibility of each pixel being HM position at each panoramic frame. Finally, the experimental results showed that the proposed DHP approach performs well in predicting HM maps of panoramic video. Thus, our work can be used to effectively model visual attention on panoramic video, having potential in practical applications, e.g., ROI-based compression of panoramic video.

Humans always perceive the world around them in a panorama, rather than the 2D plane. Therefore, it is important to model visal attention on panoramic videos for establishing human-like computer vision systems in the future. In addition, attention modelling in panoramic videos can be embedded in robotics, which need to mimic human's way in perceiving the real world. The promising future works XXX.

APPENDIX A
PROOF OF PROPOSITION 1

Example.

ACKNOWLEDGMENTS

The authors would like to thank...

REFERENCES

[1] U. Neumann, T. Pintaric, and A. Rizzo, "Immersive panoramic video," in *ACM MM*, 2000.
[2] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE TPAMI*, vol. 35, no. 1, pp. 185–207, Jan. 2013.
[3] Y. S. de la Fuente, R. Skupin, and T. Schierl, "Video processing for panoramic streaming using hevc and its scalable extensions," *Multimedia Tools and Applications*, pp. 1–29, 2016.
[4] V. R. Gaddam, M. Riegler, R. Eg, C. Griwodz, and P. Halvorsen, "Tiling in interactive panoramic video: Approaches and evaluation," *IEEE TMM*, vol. 18, no. 9, pp. 1819–1831, 2016.
[5] M. Stengel and M. Magnor, "Gaze-contingent computational displays: Boosting perceptual fidelity," *IEEE SPM*, vol. 33, no. 5, pp. 139–148, 2016.
[6] Y. Pritch, A. Rav-Acha, and S. Peleg, "Nonchronological video synopsis and indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1971–1984, 2008.
[7] Y.-C. Su, D. Jayaraman, and K. Grauman, "Pano2vid: Automatic cinematography for watching 360-degree videos," in *ACCV*, 2016.
[8] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *ICML*, 2016.
[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
[10] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE TIP*, vol. 13, no. 10, pp. 1304–1318, 2004.
[11] Y. Lin, Y. Y. Tang, B. Fang, Z. Shang, Y. Huang, and S. Wang, "A visual-attention model using earth mover's distance-based saliency measurement and nonlinear feature combination," *IEEE TPAMI*, vol. 35, no. 2, pp. 314–328, Feb. 2013.
[12] J. Zhang and S. Sclaroff, "Exploiting surroundedness for saliency detection: a boolean map approach," *IEEE TPAMI*, vol. 38, no. 5, pp. 889–902, 2016.
[13] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of ICCV*, 2015, pp. 262–270.
[14] Y. Liu, S. Zhang, M. Xu, and X. He, "Predicting salient face in multiple-face videos," in *CVPR*, 2017.
[15] T. Löwe, M. Stengel, E.-C. Förster, S. Grogorick, and M. Magnor, "Visualization and analysis of head movement and gaze data for immersive video in head-mounted displays," in *ETVIS*, 2015.

Algorithm 1 Online-DHP

Input: Panoramic video and viewer's ground-truth HM position $(x^g_{1\sim T}, y^g_{1\sim T})$, where T is the number of frame in the video;
Load offline-DHP model to initialize model with parameter vectors: $\{\theta_{\hat{\nu}}, \theta_{\hat{\pi}}, \theta_V\}$;
for $t = 1$ **to** T **do**
  for $e = 1$ **to** E **do**
    Initialize HM position: $x_1 = 0, y_1 = 0$;
    Initialize LSTM feature: $f_0 = \nvDash$;
    for $i = 1$ **to** $t-1$ **do**
      Extract FoV o_i according to (x_i, y_i);
      Obtain policy and LSTM feature: $\hat{\pi}_i, f_i = \hat{\pi}(o_i, f_{i-1}; \theta_{\hat{\pi}})$;
      Select $\hat{\alpha}_i$ according to the ϵ-greedy policy based on $\hat{\pi}_i$;
      Obtain HM magnitude: $\hat{\nu}_i = \hat{\nu}(o_i, f_{i-1}; \theta_{\hat{\nu}})$;
      According to $\hat{\alpha}_i, \hat{\nu}_i$, update (x_i, y_i) to (x_{i+1}, y_{i+1});
      Estimate reward r^{ν}_i,r^{α}_i of $\hat{\alpha}_i, \hat{\nu}_i$ according to (9), (10);
      Compute MO_i of (x_i, y_i) and (x^g_i, y^g_i);
      Store a set of experience: $\{o_i, f_{i-1}, \hat{\nu}_i, \hat{\alpha}_i, r^{\nu}_i, r^{\alpha}_i\}$;
      $i \leftarrow i + 1$;
    end for
    Update $\{\theta_{\hat{\nu}}, \theta_{\hat{\pi}}, \theta_V\}$ according to (20), (21), (22) but replace $\{\theta^{'}_{\hat{\nu}}, \theta^{'}_{\hat{\pi}}, \theta^{'}_V\}$ with $\{\theta_{\hat{\nu}}, \theta_{\hat{\pi}}, \theta_V\}$;
    $e \leftarrow e + 1$;
    $\text{MO} = \frac{\sum_i \text{MO}_i}{t-1}$;
    if $\text{MO} > th_{MO}$ **then**
      break;
    end if
  end for
  Initialize HM position: $x_1 = 0, y_1 = 0$;
  Initialize LSTM feature: $f_0 = \nvDash$;
  for $i = 1$ **to** $t-1$ **do**
    Extract FoV o_i according to (x^g_i, y^g_i);
    Obtain LSTM feature: $f_i = \hat{\pi}(o_i, f_{i-1}; \theta_{\hat{\pi}})$;
    $i \leftarrow i + 1$;
  end for
  Extract FoV o_t according to (x^g_t, y^g_t);
  Obtain policy: $\hat{\pi}_t = \hat{\pi}(o_t, f_{t-1}; \theta_{\hat{\pi}})$;
  Choose $\hat{\alpha}_t$ with the greedy policy based on $\hat{\pi}_i$;
  Obtain HM magnitude: $\hat{\nu}_t = \hat{\nu}(o_t, f_{t-1}; \theta_{\hat{\nu}})$;
  According to $\hat{\alpha}_t, \hat{\nu}_t$, update (x^g_t, y^g_t) to (x^p_{t+1}, y^p_{t+1}), which is the prediction of HM position at frame $t+1$;
  Compute MO_{t+1} of (x^p_{t+1}, y^p_{t+1}) and (x^g_{t+1}, y^g_{t+1}), which evaluates this prediction;
  $t \leftarrow t + 1$;
end for
Return: The online prediction of the viewer's HM position $(x^p_{2\sim T+1}, y^p_{2\sim T+1})$; // Every (x^p_{t+1}, y^p_{t+1}) only utilizes $(x^g_{1\sim t}, y^g_{1\sim t})$, so it is an online prediction.

[16] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360 sports video," in *CVPR*, 2017.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[18] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, no. 10, pp. 1295–1306, Jun. 2009.

[19] G. Boccignone, "Nonparametric bayesian attentive video analysis," in *International Conference on Pattern Recognition (ICPR)*, 2008.

[20] L. Zhang, M. H. Tong, and G. W. Cottrell, "Sunday: Saliency using natural statistics for dynamic analysis of scenes," in *Annual Cognitive Science Conference*, 2009, pp. 2944–2949.

[21] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE TIP*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[22] Z. Ren, S. Gao, L.-T. Chia, and D. Rajan, "Regularized feature reconstruction for spatio-temporal saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3120–3132, Aug. 2013.

[23] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient?" in *CVPR*, 2015.

[24] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang, "Learning to detect video saliency with hevc features," *IEEE TIP*, vol. 26, no. 1, pp. 369–385, 2017.

[25] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," *arXiv preprint arXiv:1510.02927*, 2015.

[26] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *European Conference on Computer Vision*. Springer, 2016, Conference Proceedings, pp. 825–841.

[27] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," *arXiv*, 2016.

[28] Ç. Bak, A. Erdem, and E. Erdem, "Two-stream convolutional networks for dynamic saliency prediction," *arXiv*, 2016.

[29] W. Wang, J. Shen, and L. Shao, "Deep learning for video saliency detection," *arXiv preprint arXiv:1702.00871*, 2017.

[30] S. Minut and S. Mahadevan, "A reinforcement learning model of selective visual attention," in *AAMAS*, 2001.

[31] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *NIPS*, 2014.

[32] M. Jaderberg, V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu, "Reinforcement learning with unsupervised auxiliary tasks," *arXiv*, 2016.

[33] Z. Wang, T. Schaul, M. Hessel, H. v. Hasselt, M. Lanctot, and N. d. Freitas, "Dueling network architectures for deep reinforcement learning," in *ICML*, 2016.

[34] J. Foote and D. Kimber, "Flycam: Practical panoramic video and automatic camera control," in *ICME*. IEEE, 2000.

[35] X. Sun, J. Foote, D. Kimber, and B. Manjunath, "Region of interest extraction and virtual camera control based on panoramic video capturing," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 981–990, 2005.

[36] Y.-C. Lin, Y.-J. Chang, H.-N. Hu, H.-T. Cheng, C.-W. Huang, and M. Sun, "Tell me where to look: Investigating ways for assisting focus in 360 video," in *ACM CHI*, 2017.

[37] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE TCSVT*, vol. 22, no. 12, pp. 1649–1668, 2013.

[38] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.

[39] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen, "A data-driven metric for comprehensive evaluation of saliency models," in *ICCV*, 2015.

[40] E. Matin, "Saccadic suppression: a review and an analysis." *Psychological bulletin*, vol. 81, no. 12, p. 899, 1974.

[41] M. J. Frederic P. Miller, Agnes F. Vandome, "Mean of circular quantities," *Mathematics*, 2010.

[42] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable mdps," in *AAAI*, 2015.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2014.

[44] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[45] I. T. Li and J. Georganas, "Multi-target multi-platform sensor registration in geodetic coordinates," in *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, vol. 1, 2002.

[46] B. Shumaker and R. Sinnott, "Virtues of the haversine," *Sky and telescope*, vol. 68, pp. 158–159, 1984.

[47] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.
[48] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, 2012.
[49] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.