IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 25, 2020, accepted January 6, 2021, date of publication January 11, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3050489

A Skill-Based Visual Attention Model for Cloud Gaming

HAMED AHMADI[1], SAMAN ZADTOOTAGHAJ[1], (Graduate Student Member, IEEE), FARHAD PAKDAMAN[1], MAHMOUD REZA HASHEMI[1], (Senior Member, IEEE), AND SHERVIN SHIRMOHAMMADI[2], (Fellow, IEEE)

[1]School of Electrical and Computer Engineering, University of Tehran, Tehran 14395-515, Iran
[2]Distributed and Collaborative Virtual Environments Research Laboratory, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Mahmoud Reza Hashemi (rhashemi@ut.ac.ir)

ABSTRACT Despite its recent advances and increasing industrial interest, cloud gaming's high bandwidth usage is still one of its major challenges. In this paper, we demonstrate how incorporating visual attention into cloud gaming helps to reduce bitrate without negatively affecting the player's quality of experience. We show that current visual attention models, which work well for normal videos, underperform in the context of cloud gaming videos. Hence, we propose our novel model, by developing a skill-based visual attention model, based on a cloud gaming dataset. First, it is demonstrated how players' attention maps are correlated with their skill levels and how this can be exploited to improve the accuracy of visual attention modeling. Then, this fact is used to cluster attention maps, according to the player's skill level. A simple yet effective method is introduced to predict players' skill levels using their performance in game. Finally, the models are incorporated into the video encoder to perceptually optimize the bitrate allocation. Incorporating the player's skill level into our model improves the accuracy of saliency maps by 14% with respect to the baseline, and 24% with respect to competing methods, in terms of Normalized Scanpath Saliency (NSS). Furthermore, we show that the maximum possible amount of video bitrate reduction depends on the player's skill level. Experimental results show 13%, 5%, and 15% reduction in video bitrate for beginner, intermediate, and expert players, respectively.

INDEX TERMS Cloud gaming, visual attention, eye tracking, video coding, perceptual video coding.

I. INTRODUCTION

Cloud Gaming (CG) was a US$1 billion industry in 2018 and is projected to grow fast to US$8 billion by 2025 [1]. CG refers to adopting the cloud computing paradigm to offer Game as a Service. In CG, the game events are captured from the players' devices and transmitted to the cloud, which processes those events, runs the game logic, renders the game scene, and streams the resulting high-quality scene in the form of video to the players [2]. CG can bring substantial benefits for both gamers and game providers, enabling ubiquitous access to gaming services while reducing cost [3]. Players no longer need to buy expensive gaming hardware at home, upgrade them every few months to avoid obsolescence, or install bulky games taking valuable local storage. In addition, the games are available anywhere and anytime, through the players' smartphone, tablet, laptop, etc. Gaming companies also benefit since they control the game cloud, facilitating game updates, piracy prevention, and cheating detection. Also, their developers need to develop for only one or at most very few target platforms, reducing the development cost. Onlive, G-cluster, and Gaikai were the pioneer companies of CG. Sony acquired Gaikai in 2012 and an essential part of Onlive in 2015 when the latter discontinued its services, enabling Sony to offer its CG service, PlayStation Now. Similarly, Nvidia offers its CG service, GeForce Now, as do many other companies such as Vortex, Playgiga, Ubitus, and others. More recently, Google announced its own CG service, Google Stadia, with much fanfare in late 2019, while Amazon revealed its own CG service, Amazon Tempo, in April 2020, with both companies investing hundreds of millions of dollars into the effort [4].

Despite CG's great promise, its performance is hindered by its high bandwidth usage and low delay requirements [2], creating a challenge to offer quality comparable to console/PC games. Even the much-hyped Google Stadia which

The associate editor coordinating the review of this manuscript and approving it for publication was Chin-Feng Lai.

has noticeable improvements over its competitors still suffers from low image quality, noticeable audio compression artifacts, stuttering on a Wi-Fi connection, and not providing true 4K resolution that players expect [5], [6], all due to the fluctuation of the available bandwidth, which forces the adaptation of the video bitrate, leading to lower quality at times.

To reduce the video bitrate without noticeable degradation in quality, one approach is to use visual attention models [7]–[10] to encode with higher-quality the regions in the scene that the player is paying more attention to, while reducing the quality of the other regions. Game Attention Model (GAM), as one such model and the first specially developed for CG, has shown prominent performance for cloud gaming [7], [8]. However, GAM works only if the game's source code is available as it produces the top-down saliency maps with the help of the game engine. This is a shortcoming, as access to the game engine's source code is not always possible. An alternative approach is to use saliency detection methods to identify visual attention. However, while these methods can be useful for natural video, we show in section II how they fail in gaming videos, as they cannot capture the logic and purpose of the game.

In this paper, we propose a new visual attention model that does not depend on the source code. Instead, it produces an offline model by learning from the players' eye-tracking data and their skill levels. Our investigations reveal that players' attention patterns are highly correlated with their skill levels. For example, experienced players, who are quite familiar with the game objects and environment, predict upcoming events in the game and focus their attention on only the screen regions which have strong ties with their success in accomplishing the game's objectives. While the visual attention patterns of all players at the same skill level are not exactly equal (due to other factors such as playing habits), they are expected to be quite similar due to common game dynamics and objectives. We further demonstrate in section IV how skill level itself can be effectively predicted using the players' score, control inputs, or their playing history.

To predict the visual attention based on skill, accurate machine learning tools such as Convolutional Neural Networks (CNN) [10], [11] can be used, but they are computationally heavy and will negatively contribute to the system's overall end-to-end delay. As mentioned earlier, the delay is another challenge in CG, hence, in this work we deploy light-weight machine learning methods.

Such a game-specific visual attention model can then determine what regions of the game the player pays more attention to, allowing the cloud gaming system to pick the best encoder configuration resulting in the most efficient bit allocation. For simplicity and without the lack of generality, in this paper, we consider the quantization parameter (QP) as the only encoder parameter that can be controlled. QP was chosen because it has the most significant impact on bitrate [12]. Assigning a higher QP to the less attended regions of the game leads to lower bitrates. Since players do not pay much attention to these regions, the overall gaming quality will be preserved [7]. Our evaluations on 18 representative game video sequences show that using our proposed model reduces the bitrate by an average of 13%, 5%, and 15% for beginner, intermediate, and expert skill levels.

To the best of our knowledge, our work is the first effort to consider the similarity of the visual attention of players at the same skill level, and use it to deduce a game attention model. The following points summarize the contributions of this paper:

- We seek alternative approaches to GAM [7] and similar eye-tracking-based methods, to perform coding optimizations according to perceptual importance.
- We propose a methodology based on offline modeling of game-specific gaze data that incorporates game-specific attention models into video encoding scheme.
- We investigate the relationship between players' skill level and game visual attention. In this regard, attention patterns are clustered for different scenarios, and players' score is used to classify them into different attention patterns.
- We incorporate the attention model into a video encoder and present a skill-based rate-controller that significantly reduces the game's bitrate, without major loss of quality.

The remainder of this paper is organized as follows. Section II reviews the related works. In Section III, we show by performance evaluation that there is a need to develop new game-specific visual attention models and we propose one such model. Section IV describes how to predict the players' skill levels, while Section V discusses incorporation of attention map into the video encoder. Section VI provides details of experimental setup and presents the experimental results. In section VII a detailed discussion is presented that illustrates various practical aspects of the proposed methodology. Finally, the paper ends with concluding remarks in section VIII.

II. RELATED WORK

Affordable eye-tracking devices are now available in the market, this is why we see several research works on efficient video compression based on gaze data for video streaming applications [9] as well as for virtual reality devices [13]. However, when the gaze data is not available, visual attention models are used instead. Incorporating visual attention to increase coding efficiency has been done in the past [7], [8], [12], [14], and many previous works have proposed different ways to computationally model human visual attention [10], [11]. It should be mentioned that a comprehensive discussion or comparison between various saliency models is beyond the scope of this paper and readers can refer to [15]–[17], and [18] for such details. In this section, we only review the models that are closest to ours and/or are being evaluated in this work.

With the recent advancement of machine learning methods, several models are proposed that use deep learning methods

such as CNNs [10] and auto-encoders [11], [19], [20]. While these models outperform the state-of-the-art models, the improvement comes with a high computational cost and inherently more delay. In this work, we use models that are light-weighted and fast to be practical for the real-time CG application.

Vig *et al.* [21] propose a fast search method, called eDN, over large-scale media to prognosticate the image saliency map. They take advantage of a hierarchical neural network model. Owing to the ample dimensionality of the parameter space, the method uses the hyper-parameter optimization strategy which is conducive to an efficient search. Thus, a simple linear classifier is generated from an optimal mixture of a multilayer feature model.

Judd's saliency model shows that a conflation of bottom-up stimulus-driven and top-down goal-driven attentions matches actual eye-movements more precisely [22]. It also combines low, middle, and high-level features and takes advantage of an SVM classifier with a linear Kernel.

Another renowned model, Boolean Map Saliency (BMS), utilizes the fact that an image can be described as a set of binary images [23]. These binary images are generated by using thresholds at various color levels. As a consequence, BMS computes the saliency map with respect to the figure-ground segregation principle. However, this model only considers bottom-up attention which makes it inappropriate for games, since a player's attention is highly bonded with his/her current task.

In Fast and Efficient Saliency detection (FES), the saliency map is produced based on the center of the image [13], using sparse sampling and kernel density estimation. Due to its limited computational complexity, this method claims to be appropriate for real-time applications as well. Additionally, it claims that people notice more the center region of each picture; hence, it introduces a feature describing the distance from the center in a Bayesian framework. However, this will decrease the accuracy of the model in games, because game elements such as enemies and collectable items are deliberately distributed over the screen and are not necessarily in the center.

Graph-Based Visual Saliency (GBVS) is proposed as a bottom-up model exploiting a substantial amount of new concepts in the literature [24]. Indeed, GBVS takes advantage of the parallel nature of graph and Markov chains over various images to form the saliency map. More precisely, it computes the edge weights with graph dissimilarity interpreted from Markov chain method. Saliency detection using region-covariances (CovSal) [25], and Itti and Koch's visual attention model (IttiKoch) [26] are other notable works.

The main drawback of the above-mentioned works is that they are not optimized to predict salient regions of game scenes. On the one hand, models that lack addressing top-down attention fail to accurately anticipate the attention patterns which are bonded with the game logic and design. On the other hand, top-down enabled attention models are all application-specific which makes them inappropriate for the game context. Other recent works such as [27]–[30] also suffer from the same shortcoming. Zadtootaghaj *et al.* [31] investigate the influencing factors that have an impact on where the players look, and categorize gaming influencing factors into three groups of user, system, and context parameters and discuss that without considering these factors, visual attention models cannot accurately predict the actual user attention patterns. In the next paragraph we demonstrate this, and we show that previous models do not perform accurately for games.

We have conducted an experiment to evaluate the performance of current models in the game context, by comparing the output of some of the models from Section II with recorded actual gaze data on a collection of game frames. Fig. 1 illustrates the results of these models on a collection of game frames. In this figure, each red dot represents where a sample player looked at while playing the game. It should be noted that since game frames are rendered based on user inputs, they differ for each player and hence, it is not possible to show the gaze location of more than one player for each game frame. In other words, each player has its own unique game frames. However, similar conclusion can be drawn from game frames of different players. The preparation of these game frames has been explained in details in the next section. As can be seen in Fig. 1, existing visual attention models fail to predict where players look at. One of the reasons is that most current eye tracking datasets comprise natural images. Moreover, they mostly have recorded gaze data while asking participants to freely view their contents. On the contrary, in a game we have to deal with computer-generated imagery. Additionally, players pay attention to the regions which are more important for the accomplishment of their current activity. In other words, the gaze location of a player during gameplay could be significantly different from merely watching the game.

One solution is to use the gaze data directly from an eye tracker. Illahi *et al.* [32] designed and implemented a system for cloud gaming with foveated graphics using a consumer grade real-time eye tracker and an open source cloud gaming platform. The gaze information was used in video compression to reduce the bitrate without a significant decrease of user experience. Zhang *et al.* [33] predict the visual attention based on touch interaction of the users and propose a tile-based optimization module to reduce the bitrate and energy. It has to be noted that their target group was not the players but the viewers of the gameplay on platforms such as Twitch.[1] Babaei *et al.* [34] demonstrate that the visual attention prediction can be improved by incorporating the game state into the model. They proposed a game attention model based on the state of the game with 17% percent improvement in accuracy.

Most of the existing perceptual solutions suffer from one or more of these shortcomings: 1) they do not consider top-down

[1] https://www.twitch.tv

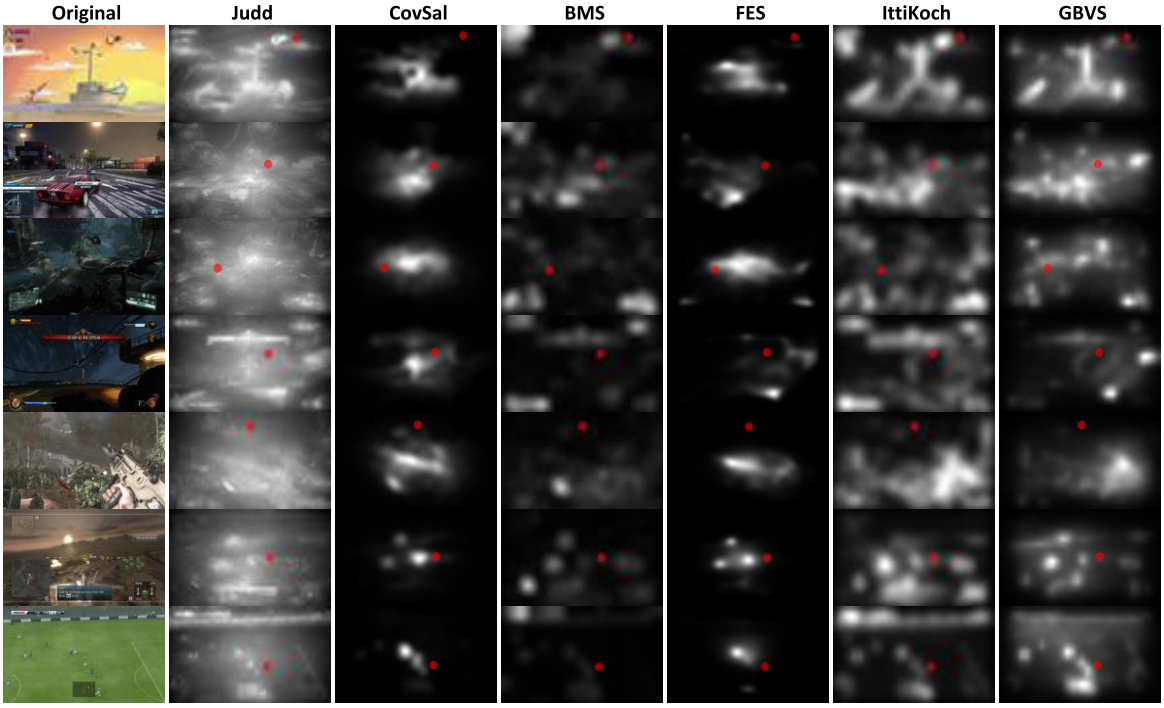

FIGURE 1. Sample game frames, with gaze locations (red dots), and the corresponding attention maps obtained for each game frame, using six different models.

information for video coding, 2) use saliency/attention models which are not tailored for game video, 3) are limited to the games with accessible code. In the following section, we present our skill-based model that estimates visual attention according to players' skill level. Unlike existing solutions, our proposed approach is tailored for cloud gaming, learns the top-down attention by offline training on a gaze dataset, and works without the need for games' code or an eye tracker.

III. VISUAL ATTENTION MODEL DEVELOPMENT

In this section, we explain the methodology by which we develop our visual attention model. Fig. 2 shows the three steps of our methodology. It also distinguishes the offline and online computation blocks of the proposed model in a deployed environment. As illustrated, all steps of building attention maps are performed completely offline. Therefore, it will not impose any delay on the server side or any power constraints on the client side, during the run-time phase. As can be seen, the first step is to conduct an eye-tracking experiment and collect gaze data from a wide range of game players covering a broad range of playing skills and habits.

The second step is to cluster the collected gaze data into several attention patterns using fuzzy C-mean clustering algorithm. The main goal is to identify similar attention patterns among groups of players, so that it can be used for smart bitrate allocation. A critical factor to consider here is the availability of the grouping metric, for assigning players to the attention clusters. For example, since it is not yet possible to lively collect eye-tracking data for every

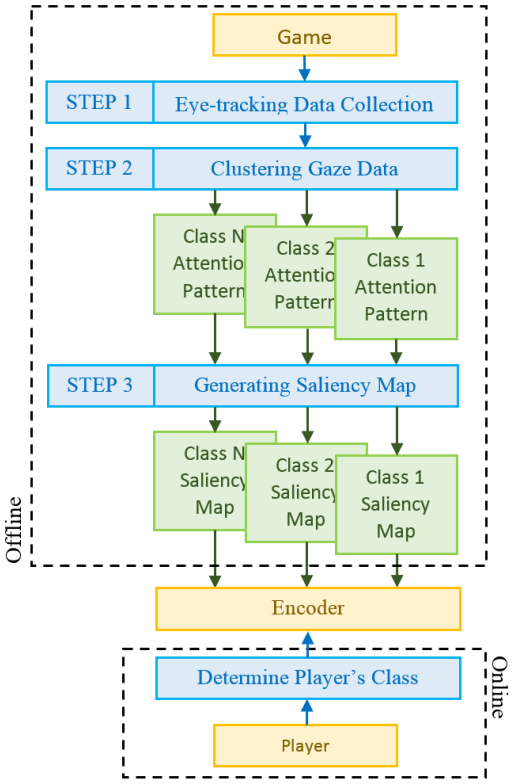

FIGURE 2. The three steps of developing the visual attention model.

cloud gaming player, the metric must not directly depend on individuals. Rather, it must rely on parameters that can be calculated or at least estimated for an individual. In this

paper, we advocate skill-based grouping of players. Skill level can not only be determined by simple measures such as analyzing their previous playing sessions and scores, the control inputs they send to the cloud gaming server, but more importantly it is observed to be significantly correlated to the player's optimum attention pattern. In this paper, we only use players' scores to predict their attention clusters. It should be noted that determining the optimal skill measure is beyond the scope of this paper, since our objective is to design a player-specific attention model to replace GAM and similar saliency methods.

The last step is to generate a saliency map for each attention cluster. Once the saliency map is ready, it is fed into the encoder, which then uses it to set the encoding parameters such as QP values such that important regions of each game frame are encoded with higher quality and other regions with lower quality. This way, bitrate is efficiently decreased without noticeably affecting the user's quality of experience. The encoding parameters can be determined as easy as fixed values or as complex as values obtained through a rate control algorithm.

In the following three subsections, we elaborate on each step of the methodology.

A. EYE-TRACKING DATA COLLECTION

Since current visual attention models are not suitable for game context, we have developed our own dataset called GSET Somi [35], the details of which can be seen in [36] and will not be repeated here, except for the following summary.

Since it is impractical for a player to keep his/her head fixed while playing a game, we did not use any chin rest, nor did we use bulky head-mount eye-trackers. We used the Tobii X2-30 Compact Eye Tracker system [37] which allows a test participant to move his/her head freely and naturally during a test session. The device's accuracy is reported to be 0.4° on average at near 60cm distance with 9-point calibration [38]. We held 135 sessions with eighty subjects, ranging 19-30 years old, twenty-three percent of whom were women. Each session comprised five steps: Introduction, Training, Calibration, Playing, and Verification. The system stored the game video sequence using a lossless compression. The duration of playing was fixed to three minutes for all participants. We ended up with 135 raw videos at 1280×720 resolution. Coupled with the collected gaze data, this is plenty of data to use for clustering and analysis, as discussed next.

B. CLUSTERING GAZE DATA

Algorithm 1 summarizes the steps required for clustering attention patterns. To do so, we first calculate each player's attention map. Each game frame is first divided into 16×16 macro-blocks. This way each frame becomes a grid of 80×45 macro-blocks. To quantize each player's gaze data in each frame, if a macro-block is gazed upon, its value is set to 1; and if a macro-block is not gazed upon, its value is considered 0. Then values of all co-located macro-blocks of game frames that a single player has watched are summed up. Next this value is divided by the total number of frames, to represent an expressive grid of 80×45 macro-blocks for each player. This process decreases the dimension of the data for each player to 3600 values (80×45). However, we still face a large number of data values for each player, which makes it hard to cluster. Hence, we use Principal Component Analysis (PCA) [39] to decrease the data dimensionality.

Algorithm 1 Clustering Algorithm for Gaze Data

```
Input: D Game videos with F frames
Output: G Cluster centroids: c_j
for each video V_s
  divide each frame into blocks of 16 × 16 pixels
  for each block B_b
    if block is gazed upon B_b = 1
    else B_b = 0
  end
  Average each block B over all frames F
end
collect D samples from all videos
find the transform matrix with P principal components
apply PCA to get P features
randomly initialize cluster memberships, μ_ij
  for each data point i and cluster j
repeat
  calculate c_j for each cluster
  update μ_ij
  calculate the objective function L
until converge
```

The number of clusters might be different from game to game, and depends on how players reach the game's objectives. For example, in racing games, most score-gaining actions require players to trace the road and react in time. On the other hand, the camera's direction is almost always such that the road is shown at the center of the screen. Therefore, the difference in attention patterns among players are less distinct than that of more complex games in which the objectives are dispersed across the screen and the camera is not fixed. Although, even in racing games, players with higher skill tend to look at the map and the rear mirror more frequently than others, which makes their attention patterns slightly different.

It should also be noted that the more distinguishable clusters are identified, the more specific attention maps could be generated and consequently, the higher the model accuracy would be. However, as the number of clusters increases, they get smaller in size and the generated map for clusters with few players would not be generalizable enough. In this work, we experimentally observe that the best number of clusters for our dataset is three. More specifically, when we cluster the data into four groups, there appears a group as small as 2 or 3 players for different numbers of PCA components. Therefore, we choose the number of clusters to be three.

For clustering, fuzzy c-mean [40] is used, which is observed to give better results than the more deterministic

methods. This can be explained by the natural fuzziness of the task (players belonging to a game cluster). To do so, a membership value, μ_{ij}, is used to indicate the degree of membership of data point x_i to cluster c_j. Using the D data points, the membership values are first initialized randomly. G cluster centroids, c_j, are calculated according to (1), where $m>1$ is used to control the degree of fuzziness.

$$C_j = \sum_{i=1}^{D} \mu_{ij}^{m} x_i \Big/ \sum_{i=1}^{D} \mu_{ij}^{m} \tag{1}$$

Next, the membership values are updated based on the centroids, using (2).

$$\mu_{ij} = 1 \Big/ \sum_{k=1}^{G} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}} \tag{2}$$

Finally, the above steps are repeated to minimize the objective function defined in (3), until the algorithm converges.

$$L = \sum_{i=1}^{D} \sum_{j=1}^{G} \mu_{ij}^{m} \|x_i - c_j\|^2 \tag{3}$$

For our dataset, we tested Algorithm 1 with different PCA-sets (different numbers of principal components). After applying PCA, we cluster players for each PCA-set where $G = 3$. Then for each of the three cluster groups in each PCA-set, the average representative frame (over all frames per each player) is calculated. Correlation average and correlation standard deviation (STD) is calculated for each cluster group between frames of the group. Clearly, the PCA-set with the largest correlation average and smallest correlation variance is the best candidate for the apt number of principal components. In our experiments, the PCA set with five principal components ($P = 5$) is the best choice, with 0.86 and 0.039 as its correlation average and STD, respectively.

C. GENERATING SALIENCY MAPS

According to the methodology presented in Fig. 2, for each cluster of attention patterns, a saliency map should be generated. To do so, we first calculate a saliency map for each game player belonging to that cluster. Then, we blend the resulting saliency maps of all players to produce the final saliency map, as depicted in Fig. 3.

In order to generate a saliency map for a single player, we first divide the screen into 80×45 blocks; each is 16×16 in size. This is, in the H.264/AVC encoder, the smallest unit on which the encoder operates. Next, for each block, we calculate the probability with which the player pays attention to that block.

Once the saliency maps of the players in a cluster are ready, we combine them into a single saliency map. This final saliency map would also include 80×45 blocks. Each block in this map is first assigned with the average of the corresponding blocks' probabilities over the cluster's saliency maps and then binarized by a threshold of 0.3. This value was determined empirically, such that on average the salient frame area is reasonably large (between 30%-40% of the frame area).

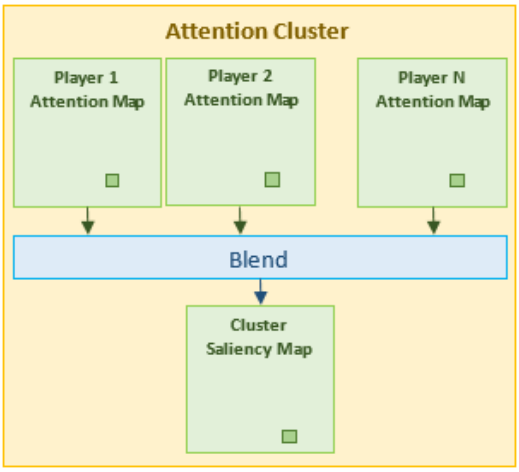

FIGURE 3. Generating a saliency map for each cluster.

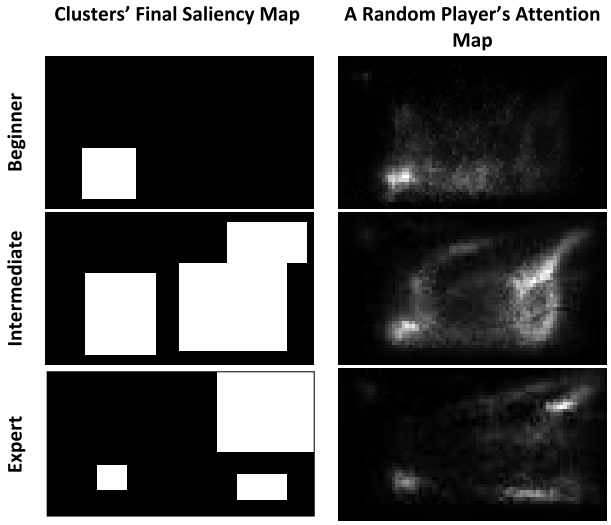

FIGURE 4. The final saliency maps of each cluster plus a random player's attention map for each skill level.

Fig. 4 shows the final saliency map for each of the three clusters. We observed that the clusters follow similar patterns for players with similar skill levels (similar experience with the game). Accordingly, we associate these clusters with three different skill levels of beginners, intermediates, and experts. In order to give readers a better sense, this figure also contains the attention map of a random player for each cluster (skill level). In the next section, we discuss how we predict each new player's skill level to assign it an attention cluster.

IV. PREDICTING A PLAYER'S ATTENTION CLUSTER

As illustrated in Fig. 2, the proposed model's output is an attention map for each identified cluster. These maps are generated offline based on the gathered eye-tracking data of each game. The next step is to determine the cluster to

which a new player belongs. Since eye-tracking devices make human-computer interaction more convenient while becoming less expensive, it would not be fictional to have them embedded in hand-held devices in the near future. If so, identifying each player's attention pattern would be easier. However, at the moment that they are not commonly available, collecting eye-tracking data for each new player and discerning his/her attention cluster is not a practical solution. Therefore, the cloud gaming system needs to estimate the player's attention cluster based on alternative parameters, such as players' skill levels. Unlike eye-tracking data, players' skill can be assessed via available data, including the user profile, previously stored on the system, and/or the control inputs s/he sends to the cloud gaming system. The next subsection details the proposed attention cluster prediction, based on skill-level.

A. SKILL-BASED PREDICTION OF ATTENTION CLUSTER

Determining the player's attention cluster, as shown in Fig. 2, is performed online. In the present work, this decision is based on the player's skill level, which is assessed based on their scores. Doing so, there will be no computational burden on the client side, although the server undertakes a slight computation to perform a look-up table retrieval and inform the encoder to incorporate the corresponding attention map.

At the start of each playing session, if there are no records available for the player, we assume the player is a beginner (zero score). As the player's score starts to build up, our model will be able to better identify the player's skill level.

According to the observations in section III, we form the hypothesis that players' skill-level causes the difference in their attention maps. In order to validate this hypothesis, we conducted an Analysis of Variance (ANOVA) [41] test. The test is performed with $\alpha = 5\%$ and gives the ratio of the "average between cluster" to the "average within-cluster" as 40.14, which is bigger than the critical value of F distribution, 3.11. Hence, it can be concluded that the score is not statistically equal among the attention clusters, proving the hypothesis. Therefore, we use the player's score to divide them into the same number of groups as we have attention clusters. We refer to these three groups as beginners, intermediates, and experts. In order to classify the players into these three groups, we find two score boundaries: the low boundary and the high boundary. Players with scores below the low boundary are considered beginners. Players with scores above the high boundary are considered experts. Other players are considered as intermediates. The choice of the boundaries will affect the accuracy of predicting the players' attention clusters, which accordingly affects the user's perceived quality and the coding gain. More specifically, using a wrong attention map has two consequences. On one hand, if the regions to which the player pays attention are considered as less important, those regions will be encoded mistakenly with lower quality and the player's perceived quality will adversely decrease. On the other hand, if the regions to which the player does not pay attention are considered as more important, those regions will be encoded unnecessarily with higher quality which diminishes the coding gain.

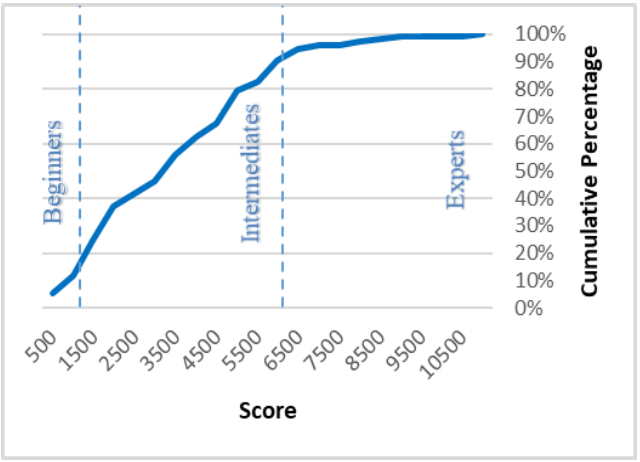

FIGURE 5. Cumulative Distribution Function of scores.

TABLE 1. Examined boundaries and their resulted accuracies.

Low Boundary	Score	Prediction Accuracy	High Boundary	Score	Prediction Accuracy
Least 5%	130	58.33%	Top 5%	7920	66.66%
Least 10%	550	66.66%	Top 10%	6240	78.57%
Least 15%	810	75.00%	**Top 15%**	**6000**	**92.06%**
Least 20%	**930**	**81.25%**	Top 20%	5780	90.47%
Least 25%	1300	80.50%	Top 25%	4820	87.30%
Least 30%	1550	79.75%	Top 30%	4510	84.12%

TABLE 2. Participants were categorized into three groups.

Group Name	SCORE RANGE
Beginner	score <= 1000
Intermediate	1000 < score <= 6000
Expert	6000 < score

In order to find the low and high score boundaries, we examined the least and top N percent of the score range, shown in Fig. 5, where $N = i \cdot \Delta N, i \in \{1, 2, \ldots, 6\}$, *and* $\Delta N = 5$. Table 1 shows the examined boundaries and their resulted accuracies. As can be seen, choosing the top score of the least 20 percent of the players ($\cong$ 1000) as the low score boundary would most accurately (81.25%) separate beginners from other players. Similarly, choosing the least score of the top 15 percent of the players (=6000) would keep experts apart from other players with an accuracy of 92.06%. Table 2 shows the score range for each group. Fig. 5 illustrates the cumulative percentage of the participants with equal to or less than a specified score. Since there are three classes, the final performance is reported as a confusion matrix in Table 3. Our test dataset consisted of twenty percent randomly-selected records, which were not used in the training phase.

TABLE 3. Confusion matrix for prediction of player's attention cluster.

		Predicted As		
		Beginner	Intermediate	Expert
Clustered As	Beginner	62.5%	37.5%	0
	Intermediate	0	97.43%	2.56%
	Expert	0	14.28%	85.71%

According to the primary diagonal of the confusion matrix in Table 3, the selected score boundaries obtain an average accuracy of 81.88%. However, we believe that this accuracy can be further improved by considering the players' previous playing sessions and/or the control inputs they send to the cloud system.

B. ATTENTION MAP UPDATE WITH GAMEPLAY

A Player's game skill does not remain constant for a person and gets improved after playing for a while. Therefore, the player's score as a performance indicator also gets improved. This fact brings us to the question of whether the player's visual attention pattern for a certain player also changes after his/her score gets increased. We investigated this question by letting a participant play the game Somi several times. It was observed that just after six repeats of playing the game, the player's performance improved significantly. Since Somi is not a complex game it is expected that the user learns quickly how to play the game. Similar to the user's performance, the user's attention pattern also changes from a pattern significantly correlated with a beginner cluster (R=0.79) to a pattern statistically correlated with the expert class (R=0.82). Fig. 6 shows the four states of a player's attention pattern over the course of six repetitions of playing the game.

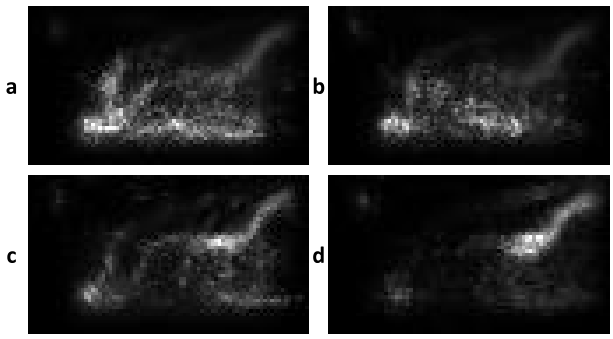

FIGURE 6. Changes in attention pattern for a player after repeated use: (a) first time playing, (b) third time, (c) forth time, (d) sixth time.

This observation once again shows the high correlation of player skill and attention pattern, which allows the cloud server to estimate the attention according to the players' skill level.

V. INCORPORATING ATTENTION MAPS INTO THE ENCODER

Once the attention maps are generated, they are ready to be incorporated into the encoder in order to select the best encoder configuration that can lead to the most perceptually efficient bit allocation [42]. In addition, it can be incorporated into other existing systems to also refine motion estimation [43] and coding unit tree decision [44] for fast encoding.

The common rationale behind all these perceptual video coding techniques is to encode region-of-interest blocks with higher quality (bitrate) and other regions with lower quality (bitrate). It should be mentioned that the coding efficiency can get improved by taking different influencing factors into account including user, content, and context factors. In this paper, the user attention pattern is taken into consideration as a user factor, to improve the coding efficiency.

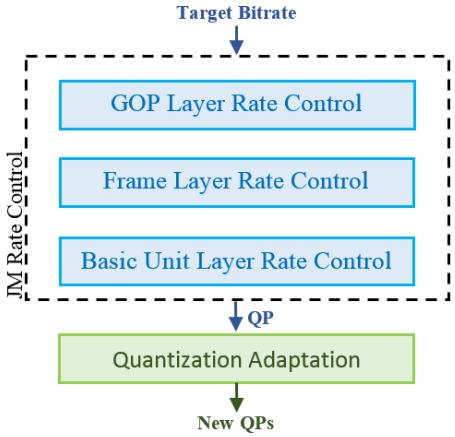

FIGURE 7. Conceptual diagram of the JM rate controller and the position of the added quantization adaptation module.

Despite newer video coding standards such as HEVC and the upcoming VVC, H.264/AVC is still one of the most commonly-used video codec [45], [46], and many cloud gaming companies use it as their basic video encoding standard [47], [48]. Hence, we present the attention map incorporation in this paper based on a rate controller similar to the one used in JM [49], which is the reference software of H.264/AVC. As shown in Fig. 7 this rate controller employs a three-layer rate control mechanism [50]. This mechanism is equipped with a quadratic rate-distortion model which uses a linear Mean of Absolute Differences (MAD) predictor to assign a QP to the basic unit. However, merely considering the encoding complexity of macroblocks without incorporating the Human Visual System (HVS) characteristics does not guarantee satisfactory user-perceived quality. Therefore, in order to assimilate the attention maps and simultaneously utilize the advantages of the original rate controller, a quantization adaptation module, as shown in Fig. 7, was added right after the rate controller. This module sets the assigned QP to the basic unit as a base QP and accordingly changes the QP of each macroblock in the

basic unit, based on that macroblock's level of importance. It should be noted that although this module is presented for JM encoder, a similar approach can be designed for any block-based video encoder.

Adapting the QP values of the macroblocks according to the visual attention map can be formulated as a global optimization problem in which the subjective quality is preserved while the bitrate is minimized [51]. According to this analysis, in the human visual characteristic based video coding, the quantization step should be inversely proportional to the attention map's value. The simulation results presented in section VI show that using this quantization adaptation leads to better subjective quality in near equal bitrates.

VI. EXPERIMENTAL RESULTS

A. EVALUATION METHODOLOGY

In this section different aspects of the proposed method are evaluated. To be able to effectively validate the performance, we first summarize the contributions of the paper here and discuss the sufficient experiments and metrics to evaluate each contribution. Then, in next subsections we discuss the details of the experiments.

The main novelties discussed in this paper are (1) a methodology to model the visual attention pattern of cloud gaming videos, (2) establishing the relationship between players' skill level and game visual attention, and (3) encoding rate adaptation based on attention patterns. Accordingly, we have conducted experiments to evaluate these aspects.

For case (1), section B measures the accuracy of the proposed models and compares the accuracy with several competing methods. To further validate the accuracy of the proposed models according to the human perception, we have also measured the Normalized Scanpath Saliency (NSS) [52] and compared it with a competing method. NSS is defined as the response value at the current eye position, in a model's predicted gaze density map that has been normalized to have zero mean and unit standard deviation. Hence, it can be used as a robust measure of accuracy in attention maps.

For case (2), we have already presented sufficient experiments in section IV that establish the strong correlation between players' skill-levels and visual attentions. Specifically, it has been demonstrated how visual attention changes gradually as a player's skill develops (Fig. 6). Moreover, the effectiveness of assigning attention patterns to players based on their scores (as a measure of skill) has been evaluated and reported in Tables 1 and 3.

For case (3), we have conducted extensive experiments including integrating attention maps into the video encoder, and adapting bitrate according to visual importance. The performance of this scheme is compared to the baseline encoding. These experiments are repeated for constant bitrate coding, as well as constant quality coding, to measure the performance in different encoding strategies. Both bitrate and quality are measured for comparisons. While the encoding gain is measured by calculating the bitrate reduction, the quality is measured using Eye-tracking Weighted Peak Signal-to-Noise Ratio (EWPSNR) [51]. EWPSNR weights the distortions according to their distance and angle with respect to the subject's fixation points. As it measures the distortions considering the players' fixation points, it is a very common alternative for subjective evaluation [53]–[56]. Moreover, EWPSNR is known to be a fair perceptual objective metric since it is independent of visual attention models.

Next subsections detail the experiments and discuss the results.

B. VISUAL ATTENTION MODEL EVALUATION

In order to evaluate the proposed model, we measure how accurately it predicts the gaze locations. To do so, we calculate the number of times that gaze locations fall onto salient regions of game scenes. However, comparing its performance with that of the current state-of-the-art visual attention models needs a consideration. More specifically, in order to binarize the output of these models, we need to apply a threshold to them. This threshold affects the salient area of the final maps. For a given model, the smaller the threshold is, the smaller the salient regions will be. So will be the model's accuracy in predicting gaze locations. Therefore, a fair comparison mechanism should be first established. To this end, for each competing model, we experimentally find the threshold at which the average area of the model's output saliency maps is equal or greater than that of the proposed model.

Table 4 shows the accuracy of the proposed visual attention prediction models for each skill level. As can be observed, the proposed model outperforms the previous ones in the game context. To clarify the comparison procedure, we explain how to calculate GBVS's accuracy for beginners as an example. For the other models or skill levels, the accuracy is calculated similarly. For each beginner, we first apply GBVS on all frames of the player's game video. Second, we binarize the resulting saliency maps with a threshold value. Then, we calculate the average percentage of salient regions within the frame area. Next, we calculate the average of these percentages among all beginner players and call it Salient Area Percentage (SAP). This procedure is repeated for different threshold values starting from 0.1 to 1.0 until the SAP is equal or greater than that of the proposed method (i.e. 6.67%). For GBVS, it happens to be 0.5. This threshold is considered as a good comparison point, because its SAP is close to the proposed model's SAP which is an indication that if we encode the videos, once with the help of GBVS and another time with the help of the proposed model, the encoded videos will have approximately equal bitrates. After determining the appropriate threshold, we count the number of times that gaze locations fall onto the salient regions of the thresholded saliency maps. Finally, the model's accuracy is calculated by dividing this number by the total number of gaze locations. According to Table 4, for beginners, applying a threshold of 0.5 on GBVS's output

TABLE 4. Accuracy of the models in predicting gaze locations.

BEGINNERS			
Model	THRESHOLD	SALIENT AREA (%)	ACCURACY (%)
Proposed	0.3	**6.67**	**50.05**
GBVS	0.5	8.56	13.34
ITTI	0.4	7.32	14.64
BMS	0.8	7.68	10.74
FES	0.7	8.36	5.91
JUDD	0.3	13.3	29.49
CovSal	0.6	7.60	3.86
INTERMEDIATES			
Model	THRESHOLD	SALIENT AREA (%)	ACCURACY (%)
Proposed	0.3	**45.11**	**96.68**
GBVS	0.9	58.94	96.11
ITTI	0.9	63.76	86.63
BMS	1.0	94.33	96.67
FES	1.0	49.99	88.53
JUDD	0.5	38.20	72.36
CovSal	0.98	50.59	93.98
EXPERTS			
Model	THRESHOLD	SALIENT AREA (%)	ACCURACY (%)
Proposed	0.3	**25.47**	**73.36**
GBVS	0.8	40.07	42.15
ITTI	0.7	29.81	40.45
BMS	0.9	35.65	55.03
FES	1.0	49.81	54.21
JUDD	0.5	36.76	37.68
CovSal	0.9	25.92	15.47

saliency maps will result in an SAP of 8.56%. Although this is bigger than 6.67%, GBVS's accuracy is significantly less than the accuracy of the proposed model.

The accuracy of the proposed method for beginner, intermediate, and expert players is 50.05%, 96.68%, and 73.36% respectively which outperform the competing methods. It should be noted that the threshold of binarization of saliency maps in our model is kept the same for all three groups of players, i.e. 0.3, which results in different salient area percentages for each player group. Therefore, higher accuracy is achieved for intermediate players which obtained a wider salient area and lower accuracy for beginners which obtained a more limited salient area. In other words, these results do not imply that the proposed model performs better for intermediate compared to other groups, and this difference is due to different attention patterns of different skill levels.

Comparing with the competing methods, it is observed that the proposed method achieves the best results among all. For intermediates, the proposed method gains a 96.68% accuracy with an SAP of 45.11%. Some competing methods also gain an accuracy close to this, however, with a much larger SAP. For instance, BMS gains a 96.67% accuracy with an SAP of 94.33%. This indicates that although both methods gain excellent accuracy, BMS produces an attention map which is too large to be used to save bitrate effectively.

For beginners and experts, as mentioned above, the attention maps are more concentrated and hence more challenging. For beginners the proposed method gains an impressive accuracy of 50.05% with an SAP as small as 6.67%. The next best result, 29.49%, belongs to JUDD with an SAP of 13.3%, which is almost twice as large. For experts, the proposed method gains 73.36 with an SAP of 25.47%. The next best result belongs to BMS with 55.03% accuracy with an SAP of 35.64. These results show that not only the proposed method provides a high accuracy attention map, but the map is also small enough, such that it can be used to optimize the video encoding to save bitrate.

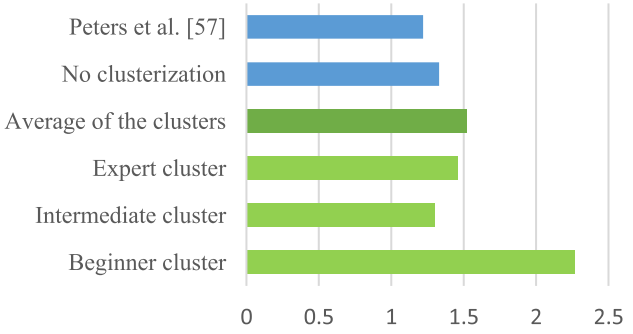

FIGURE 8. NSS of different scenario.

To further evaluate our model, we measured the Normalized Scanpath Saliency (NSS) for each skill level's attention and compare with the attention model proposed in [57]. As can be seen in Fig. 8, the NSS of the proposed model is 24.59% higher on average. Furthermore, this figure shows that clustering attention patterns would result in a 14.28% increase in NSS rather than having no clustering. Compared to the other two clusters, the lower NSS of the Intermediate cluster can be characterized as the higher variety in attention patterns among intermediate game players. Unlike experts, who focus on the most score-gaining parts of the screen, and beginners who unnecessarily spend much time on limited parts of the screen, intermediates are skilled and curious enough to explore different elements in the game. Hence, their looking at more and wider regions of the screen has resulted in a larger salient area in this cluster's attention map. A larger salient area consequently increases the likelihood of false-positive occurrences, which are penalized by NSS and lead to a lower value of NSS.

It should be noted that when a visual attention model is incorporated into a video encoder, its accuracy also affects the perceived quality of the encoded video. If an attended region were mispredicted as unattended, the attended region would be encoded with low quality, which will be perceived by viewers, and affects the visual quality.

C. EVALUATING THE RATE ADAPTATION MODULE

This section evaluates the performance of rate adaptation based on attention maps. The JM V18.4 is used for the experiments. In order to measure the impact of utilizing

the generated attention maps on the coding gain, we conducted two sets of trials. First, the JM rate controller is turned off and the QP values are experimentally changed according to the attention map. Second, we compare the performance of the original JM rate controller and the visual attention-based quantization adaptation module, described in section V. In both sets, we encode the collected game videos, using the reference software of H.264/AVC. In order to put macroblocks of the same importance into separate slices, we activate the Flexible Macroblock Ordering (FMO) tool which is one of the several error resilience tools defined in the Baseline profile of this standard. Since our video games have been recorded with 720p resolution, the suitable level of the Baseline profile is 3.1. We encode each video sequence with the help of its corresponding saliency map. Finally, we compare the bitrate and objective quality of the decoded versions of those video sequences. In all cases, we choose a Group of Pictures (GOP) size of 15. Table 5 summarizes the encoding parameters that were used in our simulations.

TABLE 5. H.264/Avc encoding parameters used in our evaluations.

Parameter Name	PARAMETER VALUE
Profile	Baseline
Level	3.1
Number of Reference Frames	1
Motion Estimation Scheme	EPZS
Search Range	32
RD-optimized mode decision	ON

1) RATE-CONTROL-OFF EXPERIMENTS

We encode the game video sequences, collected during the eye-tracking sessions, in two ways: 1) with a single QP value and 2) with multiple QP values. The game video sequences, are first encoded without attention maps and hence using a single-QP value. In order for the comparison to be fair, we asked several experts in video encoding to choose the maximum QP value for which they cannot detect any distortion in the video frames. This maximum QP value turned out to be 26 and the average PSNR of the references for this QP value is 45.32 dB.

The reference video sequences are once again encoded, this time with the help of attention maps. In this case, each sequence is encoded with two QP values corresponding to the priority regions. We investigate all possible ordered pairs of the form (26, 26 + 2X) where $1 <= X <= 3$.

Table 6 shows EWPSNR and bitrate of eighteen sample video sequences that have been encoded once with a single-QP value of 26 and other times with the QP sets mentioned above.

Table 7 shows the bitrate reduction percentage achieved using multiple QP values compared to the constant QP of 26. It should be noted that the QP set needs to be chosen based on available bandwidth. In addition, the QP values should not be too much apart since a significant difference between regions catches players' attention patterns and significantly affects the gaming experience.

It can be observed that as the QP difference between salient and non-salient areas increases, the bitrate reduction improves while the quality decreases. Overall, a four-step difference (the pair of 26 and 30) is observed to gain reasonable bandwidth reduction as well as fair quality. For beginners, intermediates, and experts, this configuration reduces the bitrate by 29.58%, 10.87%, and 26.25%, while losing only 1.7 dB, 0.4 dB, and 1.42 dB compared to the baseline encoding, respectively. The average qualities for these cases are 42.41 dB, 43.49 dB, and 44.20 dB, which are quite high. Furthermore, it is observed that a smaller bitrate saving is archived for intermediate players. This is due to the wider saliency maps of intermediate group, as discussed in section VI.B.

2) RATE-CONTROL-ON EXPERIMENTS

In order to compare the performance of the JM rate controller and the visual attention-based quantization adaptation module, we measure EWPSNR of the decoded game video sequences in both cases. Fig. 9 shows an example of EWPSNR over one hundred and twenty consecutive frames of the game video sequence belonging to an intermediate game player. As can be seen from the figure, the quality of the quantization adaptation module is in general better than the JM rate controller. However, in a few frames, the quality difference is negative, which is not significant and does not affect the overall gain in quality. The reason could be the misprediction of the attention model. Table 8 shows the average EWPSNR and bitrate of the decoded video sequences for the JM rate controller and the quantization adaptation module. First, it is observed that the rate controller with the quantization adaptation module gains a similar bitrate as the baseline encoder. This is important as it means that the quantization adaptation does not interfere with the rate controller. Second, it can be observed that the proposed method achieves a better quality compared to the baseline encoding, with the same bitrate. For beginners, intermediates, and experts, the quantization adaptation module gains a 0.65 dB, 1.27 dB, and 1.97 dB increase in EWPSNR, respectively.

VII. DISCUSSION AND FUTURE WORK

The proposed methodology in this paper, to develop a skill-based visual attention model, is general enough to be utilized for any video game. However, the implementation of its steps might differ from game to game. In general, there are several considerations that should be taken into account.

In this paper, we showed that the current conventional video attention models would not perform well for video games. One reason behind this low performance is the high importance of top-down attention of players. In fact, based on the game rules and player experience in the game, players learn to look at special parts of the game-scene which might

TABLE 6. EWPSNR and bitrate of the video sequences for different QP values EWPSNR and Bitrate have been reported in dB and Mbps, respectively Bi, Ii, and Ei are beginner, intermediate, and expert players, respectively.

	B1		B2		B3		B4		B5		B6	
Scenario	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate
QP (26)	43.33	1.94	43.95	1.90	44.19	2.13	45.17	2.07	44.47	1.77	43.85	2.09
QPs (28 , 26)	41.93	1.64	43.54	1.61	43.34	1.78	44.64	1.74	43.83	1.49	42.75	1.77
QPs (30 , 26)	40.41	1.37	43.11	1.35	42.31	1.49	44.05	1.45	43.12	1.24	41.50	1.48
QPs (32 , 26)	38.93	1.16	42.60	1.14	41.28	1.24	43.35	1.23	42.32	1.05	40.21	1.25

	I1		I2		I3		I4		I5		I6	
Scenario	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate
QP (26)	44.22	2.37	45.35	4.01	42.81	1.80	43.23	1.90	43.76	2.12	44.00	2.09
QPs (28 , 26)	44.12	2.27	45.11	3.68	42.51	1.72	43.07	1.81	43.70	2.00	43.8	2.00
QPs (30 , 26)	43.97	2.17	44.82	3.35	42.14	1.63	42.87	1.71	43.60	1.87	43.58	1.90
QPs (32 , 26)	43.80	2.08	44.51	3.09	41.71	1.56	42.62	1.63	43.48	1.77	43.33	1.81

	E1		E2		E3		E4		E5		E6	
Scenario	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate	EW PSNR	Bitrate
QP (26)	45.16	2.79	46.31	2.21	47.16	2.00	44.02	1.89	44.79	1.95	46.30	2.76
QPs (28 , 26)	44.11	2.34	45.32	1.93	46.47	1.72	43.80	1.65	44.51	1.68	45.68	2.36
QPs (30 , 26)	42.98	1.95	44.11	1.69	45.59	1.46	43.48	1.44	44.16	1.46	44.93	1.99
QPs (32 , 26)	41.87	1.64	42.75	1.49	44.70	1.26	43.16	1.27	43.77	1.27	44.22	1.70

TABLE 7. Bitrate reduction percentage of multi-qp scenarios Bi, Ii, and Ei are beginner, intermediate, and expert players, respectively.

	B1	B2	B3	B4	B5	B6	Average (%)
QPs (28 , 26)	15.46	15.26	16.43	15.94	15.82	15.31	15.71
QPs (30 , 26)	29.38	28.95	30.05	29.95	29.94	29.19	29.58
QPs (32 , 26)	40.21	40	41.78	40.58	40.68	40.19	40.57

	I1	I2	I3	I4	I5	I6	Average (%)
QPs (28 , 26)	4.22	8.23	4.44	4.74	5.66	4.31	5.27
QPs (30 , 26)	8.44	16.46	9.44	10	11.79	9.09	10.87
QPs (32 , 26)	12.24	22.94	13.33	14.21	16.51	13.4	15.44

	E1	E2	E3	E4	E5	E6	Average (%)
QPs (28 , 26)	16.13	12.67	14	12.7	13.85	14.49	13.97
QPs (30 , 26)	30.11	23.53	27	23.81	25.13	27.9	26.25
QPs (32 , 26)	41.22	32.58	37	32.8	34.87	38.41	36.15

TABLE 8. Average EWPSNR for JM rate controller and quantization adaptation module EWPSNR and bitrate have been reported in dB and Mbps, respectively.

	JM Rate Controller		Quantization Adaptation		Difference	
Scenario	EWPSNR	Bitrate	EWPSNR	Bitrate	EWPSNR	Bitrate
Beginner	40.19	1.91	40.84	1.92	+0.65	-0.01
Intermediate	42.74	2.37	44.01	2.32	+1.27	+0.05
Expert	41.63	1.71	43.6	1.71	+1.97	0.00

be considered simple from a video-saliency point of view but lead to a better score. This causes the state-of-the-art attention models to perform poorly on game video. We proposed a new framework to develop game attention models that takes into account the user skill and depends on the strategy of the player in the video games. This work shows how to use the gaze information of users to build such a customized attention model.

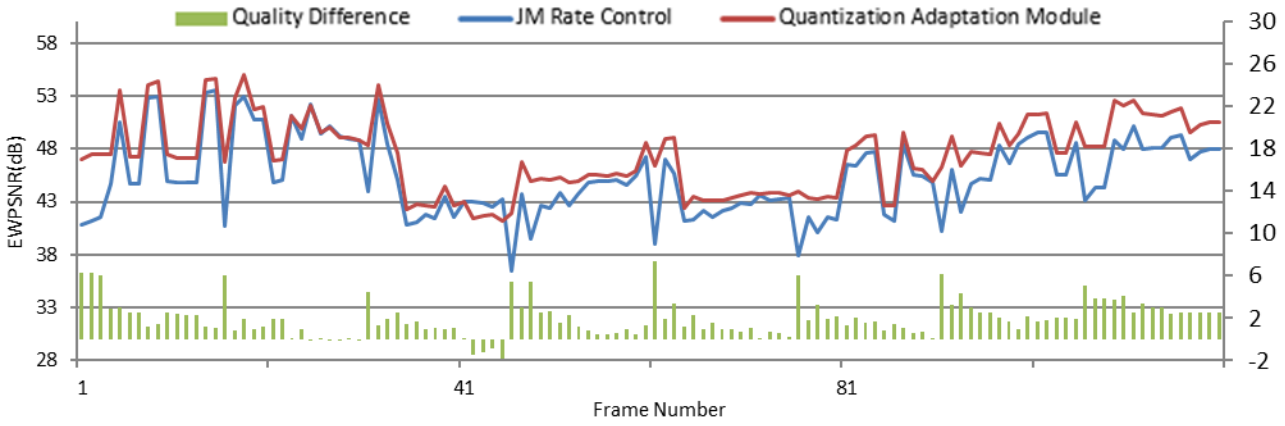

FIGURE 9. EWPSNR of 120 consecutive frames of a sample intermediate player.

In this paper, we focused on one video game, however, the proposed methodology can be applied to any other game by following its steps, starting with the data collection. The collection of eye-tracking data costs both time and money. Thus, the gain achieved from incorporating the model must be economically justifiable. On the one hand, video games with more scenes and situations require a larger amount of eye-tracking data, consequently increasing the costs. On the other hand, reducing the required bitrate for that video game would help in increasing the number of potential customers and consequently improve revenue. Therefore, before utilizing the model, service providers should conduct a cost-benefit analysis. For example, it takes seven hours on average to play the main story of Call of Duty: Black Ops II [58], one of the most popular first-person shooter video games [59]. It means that collecting 200 sessions for this video game would roughly take 20 days, assuming we have 10 eye-tracking devices and one session running on each device per day. Since the time between two consecutive releases of a modern video game is at least one year and each release usually sells tens of million dollars [60], such costs can be easily recovered. The dataset is required to include as many game situations as possible. Here, by the game situation, we mean a combination of game objects, states, and scenes. The more situations are covered in the dataset, the more accurate the model would be. If these game situations are distinguishable enough, the clustering step in the methodology allows our technique to identify and address them separately. Otherwise, their effects will be handled in the next step (generating saliency maps) where we calculate the probability with which players pay attention to different regions of the game video. If a distinguishable long-lasting game situation is not covered in the dataset, it would adversely affect the accuracy of the model when the player enters that situation. However, this problem is common among all learning algorithms. The more representative the dataset is, the more accurate and fitter the model would be.

It should be noted that this paper proposes a methodology for cloud gaming and not a rigid solution. This means that different steps of the methodology can be followed to achieve visual attention models for a new game. For instance, a feature reduction scheme is suggested after data collection. We show that the best feature reduction can be decided by repeating Algorithm 1 for different numbers of PCA components and deciding based on highest achieved correlation. However, details of these steps may be changed or fine-tuned according to the game or service requirements. For instance, a clustering step should be done to group attention patterns into different clusters. In our example, we use three clusters, which also sets the granularity for skill-levels. However, this granularity depends on the complexity of the game and may be set differently. Regardless of the number of clusters, the main point of this paper, which is to use skill-levels for clustering attention patterns, holds for all games and can benefit the cloud gaming service.

It should also be noted that in order to gain higher NSS, the proposed attention model should be generated for each game separately. The reason is that each game has its own unique game logic and design. These two parameters define the game objectives which direct players' top-down goal-driven attention. Therefore, if a dataset is collected from game A, it can not necessarily be used for game B.

The model assumes that players with the same skill level have quite similar attention patterns. There are, however, other factors such as the style of playing which would also drive players' attention. But, since fulfilling the game objectives dictates specific actions and considerations to game players, their attentions' degree of freedom is restricted. It means that although attention patterns are not exactly equal among players with the same skill level, they are similar enough to be grouped into one category. The second step of our proposed methodology does exactly that: clustering these groups of patterns from players with the same skill level, even though the patterns are not exactly the same. This results in a more finely tuned skill-based visual attention model, as proven by our experimental results. In order to further investigate the impact of skill and other factors on attention patterns in different game genres and under

various circumstances, researchers need to collect much more eye-tracking data which is a burdensome task. We have organized and annotated the current version of our dataset and made it publicly available [35]. As our future work, we plan to continue our eye-tracking data collection and expand our dataset. This will boost academic research pertaining to game-specific visual attention models and also provide a common basis for comparing such models. For example, currently, we have started the process of eye-tracking collection for a platform video game, titled ''Rayman Legends'' [61]. Although in the early stages, our preliminary tests show that the impact of players' skill on attention patterns can be verified in this game too.

One of the practical applications of the proposed methodology for future work is to extend it to be used in Head-Mounted Displays (HMD) where the gaze data can be collected using head movement or embedded eye tracker. Such a scheme can be used for perceptual enhancement of 360° video and virtual reality-based games [62], [63].

Another point to discuss is how to accurately recognize the player's skill level or more specifically their corresponding attention cluster. In this paper, we used the player's score to predict their skill level, because score and skill are highly correlated. Studies have shown that game score is influenced by factors such as player's skill, network latency, and network jitter [64]. Although the influence of each of these factors depends on the game genre, one solid conclusion is that skill is the factor that influences the game score the most, across a great majority of genres, possibly all [64]. Since in our tests not only delay and jitter but also hardware and software (such as processing power, graphics fidelity, input equipment, etc.) are the same for all players, we can confidently depend on player's skill as the only factor affecting the score. Therefore, score can be used as a relevant feature to predict players' attention patterns. However, it should be noted that the score cannot fully represent a participant's skill, especially for beginner players. Clearly, a beginner player who shoots constantly may get a high score which does not represent his/her skill level. Hence, players are mistakenly classified into intermediate class, as is observed in Table 3. Therefore, combining the score with other relevant objective metrics such as ''action per minute'' and ''input pattern'' can definitely improve the classification of players. Moreover, each game has its own content and requires its own considerations. For example, in some video games, players do not receive explicit scores. However, their skill levels can still be predicted based on other in-game parameters such as gained experience points, power-ups, and/or equipment. Furthermore, players themselves generate distinctive information on their skill levels. As our future work, we plan to investigate the possibility of utilizing a player's control inputs in the prediction of their attention clusters.

Finally, the methodology presented in this paper was evaluated on H.264/AVC, as it is one of the most commonly used video coding standards. Despite this, the same methodology can be adapted for the newer standards, such as HEVC, VVC, or AV1. All these codecs use block-based coding, which can be used in a similar way as this paper for adjusting the quality. In fact, the newer codecs such as HEVC and VVC, use a more flexible coding block structure, which allows a more detailed adjustment of the quality based on attention maps. This can be used in future works both for improving the coding performance and having a smoother change of quality throughout different saliency regions. However, these modern codecs are more computationally complex and have longer encoding delays [65]. Given that delay has a great impact on players' quality of experience, this should be mitigated using efficient video compression techniques such as fast mode decision [66], [67], fast motion estimation [43], and delay-aware video compression [44], [68].

VIII. CONCLUSION

In this paper, we showed that current visual attention models underperform in the game-context. Using the results of our previously conducted experiment on where game players tend to look in a game scene, we have developed a more efficient game-specific visual attention model. The proposed model predicts the salient regions of game scenes taking the player's skill level into account. Our evaluations show that the proposed model performs more accurately in the game context, compared to the current conventional visual attention models. In this paper, we also utilized the model with an H.264/AVC encoder and concluded that it is possible to reduce the required bitrate of cloud gaming by an average percentage of 13, 5, and 15 for beginner, intermediate, and expert skill levels, respectively.

REFERENCES

[1] A. Bhutani and P. Wadhwani, ''Cloud gaming market size market share & forecast 2019–2025,'' Global Market Insights, Pune, India, Tech. Rep. ID GMI2368, 2019.

[2] S. Shirmohammadi, M. Abdalla, D. T. Ahmed, K.-T. Chen, Y. Lu, and A. Snyatkov, ''Introduction to the special section on visual computing in the cloud: Cloud gaming and virtualization,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1955–1959, Dec. 2015.

[3] R. Shea, J. Liu, E. C.-H. Ngai, and Y. Cui, ''Cloud gaming: Architecture and performance,'' *IEEE Netw.*, vol. 27, no. 4, pp. 16–21, 2013.

[4] *Amazon Pushes Into Making Video Games, not Just Streaming their Play*. Accessed: May 4, 2020. [Online]. Available: https://www.nytimes.com/2020/04/02/technology/amazon-making-video-games.html

[5] A. Goslin. *Streaming Assassin's Creed Odyssey in Google Chrome is Surprisingly Great*. Accessed: May 4, 2020. [Online]. Available: https://www.polygon.com/2018/10/8/17953130/project-stream-game-streaming-assassins-creed-odyssey

[6] S. Etienne. *Google's Project Stream is a Working Preview of the Future of Game Streaming*. Accessed: May 4, 2020. [Online]. Available: https://www.theverge.com/2018/10/8/17950998/google-project-stream-gaming-assasins-creed-odyssey-first-impression

[7] H. Ahmadi, S. Z. Tootaghaj, M. R. Hashemi, and S. Shirmohammadi, ''A game attention model for efficient bit rate allocation in cloud gaming,'' *Multimedia Syst.*, vol. 20, no. 5, pp. 485–501, Oct. 2014.

[8] H. Ahmadi, S. Khoshnood, M. R. Hashemi, and S. Shirmohammadi, ''Efficient bitrate reduction using a game attention model in cloud gaming,'' in *Proc. Haptic Audio Vis. Environ. Games*, 2013, pp. 103–108.

[9] J. Ryoo, K. Yun, D. Samaras, S. R. Das, and G. Zelinsky, ''Design and evaluation of a foveated video streaming service for commodity client devices,'' in *Proc. 7th Int. Conf. Multimedia Syst. (MMSys)*, May 2016, pp. 1–11.

[10] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.
[11] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, Oct. 2018.
[12] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.
[13] H. R. Tavakoli, E. Rahtu, and J. Heikkila, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Image Analysis*. Berlin, Germany: Springer, 2011, pp. 666–675.
[14] M. Rerabek, H. Nemoto, J.-S. Lee, and T. Ebrahimi, "Audiovisual focus of attention and its application to ultra high definition video compression," *Proc. SPIE*, vol. 9014, Feb. 2014, Art. no. 901407.
[15] H. Liang, R. Liang, and G. Sun, "Looking into saliency model via space-time visualization," *IEEE Trans. Multimedia*, vol. 18, no. 11, pp. 2271–2281, Nov. 2016.
[16] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.
[17] Y. Ji, H. Zhang, Z. Zhang, and M. Liu, "CNN-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances," *Inf. Sci.*, vol. 546, pp. 835–857, Feb. 2021.
[18] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 19, 2019, doi: 10.1109/TPAMI.2019.2935715.
[19] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder-decoder network for visual saliency prediction," Feb. 2019, *arXiv:1902.06634*. [Online]. Available: http://arxiv.org/abs/1902.06634
[20] Y. Ji, H. Zhang, Z. Jie, L. Ma, and Q. M. J. Wu, "CASNet: A cross-attention siamese network for video salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 21, 2020, doi: 10.1109/TNNLS.2020.3007534.
[21] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.
[22] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.
[23] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 153–160.
[24] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.
[25] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.*, vol. 13, no. 4, p. 11, Mar. 2013.
[26] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[27] C. Shen, X. Huang, and Q. Zhao, "Predicting eye fixations on webpage with an ensemble of early features and high-level representations from deep network," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2084–2093, Nov. 2015.
[28] J. Lei, B. Wang, Y. Fang, W. Lin, P. L. Callet, N. Ling, and C. Hou, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, Sep. 2016.
[29] Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Human visual system-based saliency detection for high dynamic range content," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 549–562, Apr. 2016.
[30] Z. Wang, D. Xiang, S. Hou, and F. Wu, "Background-driven salient object detection," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 750–762, Apr. 2017.
[31] S. Zadtootaghaj, S. Schmidt, H. Ahmadi, and S. Möller, "Towards improving visual attention models using influencing factors in a video gaming context," in *Proc. 15th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Jun. 2017, pp. 1–3.
[32] G. K. Illahi, T. V. Gemert, M. Siekkinen, E. Masala, A. Oulasvirta, and A. Ylä-Jääski, "Cloud gaming with foveated video encoding," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1, pp. 1–24, Apr. 2020.
[33] C. Zhang, Q. He, J. Liu, and Z. Wang, "Exploring viewer gazing patterns for touch-based mobile gamecasting," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2333–2344, Oct. 2017.
[34] E. Babaei, M. R. Hashemi, and S. Shirmohammadi, "A state-based game attention model for cloud gaming," in *Proc. 15th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Jun. 2017, pp. 1–3.
[35] H. Ahmadi, S. Z. Tootaghaj, S. Mowlaei, M. R. Hashemi, and S. Shirmohammadi, "GSET Somi: A game-specific eye tracking dataset for Somi," *IEEE Dataport*, 2020. Accessed: Oct. 26, 2020, doi: 10.21227/bvt7-3b15.
[36] H. Ahmadi, S. Z. Tootaghaj, S. Mowlaei, M. R. Hashemi, and S. Shirmohammadi, "GSET Somi: A game-specific eye tracking dataset for Somi," in *Proc. Int. Conf. Multimedia Syst.*, 2016, pp. 1–6.
[37] *Tobii X2 Eye Tracker Portable Lab*. Accessed: Apr. 16, 2020. [Online]. Available: https://www.tobiipro.com/product-listing/tobii-pro-x2-30/
[38] *Tobii X2-30 Eye Tracker Accuracy and Precision Test Report*. Accessed: Apr. 16, 2020. [Online]. Available: https://www.tobiipro.com/siteassets/tobii-pro/accuracy-and-precision-tests/tobii-x2-30-eye-tracker-accuracy-and-precision-test-report.pdf
[39] I. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Springer, 2002, pp. 167–198.
[40] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York, NY, USA: Springer, 1981.
[41] A. Field, *Discovering Statistics Using IBM SPSS Statistics*. Newbury Park, CA, USA: Sage, 2013.
[42] J.-S. Lee and T. Ebrahimi, "Perceptual video compression: A survey," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 684–697, Oct. 2012.
[43] F. Pakdaman, M. R. Hashemi, and M. Ghanbari, "A low complexity and computationally scalable fast motion estimation algorithm for HEVC," *Multimedia Tools Appl.*, vol. 79, nos. 17–18, pp. 11639–11666, Jan. 2020.
[44] X. Deng, M. Xu, L. Jiang, X. Sun, and Z. Wang, "Subjective-driven complexity control approach for HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 91–106, Jan. 2016.
[45] *Video Encoding: The Definitive Guide*. Accessed: May 4, 2020. [Online]. Available: https://www.dacast.com/blog/what-is-video-encoding/
[46] *Video Codecs and Encoding: Everything You Should Know*. Accessed: May 4, 2020. [Online]. Available: https://www.wowza.com/blog/video-codecs-encoding
[47] *Nvidia NVENC Outperforms AMD VCE on H.264 Encoding Latency in Parsec Co-Play Sessions and How it Impacts Overall Lag*. Accessed: Apr. 30, 2020. [Online]. Available: https://blog.parsecgaming.com/nvidia-nvenc-outperforms-amd-vce-on-h-264-encoding-latency-in-parsec-co-op-sessions-713b9e1e048a
[48] *Steamworks Documentation, Steam Video*. Accessed: Apr. 30, 2020. [Online]. Available: https://partner.steamgames.com/doc/features/streaming_video
[49] *H.264/AVC Reference Software*. Accessed: Apr. 14, 2020. [Online]. Available: http://iphome.hhi.de/suehring/tml
[50] W. Gao and F. Pan, *Adaptive Rate Control With HRD Consideration*, document: JVT-H014, 2003.
[51] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, Jan. 2011.
[52] R. J. Peters, A. Iyer, C. Koch, and L. Itti, "Components of bottom-up gaze allocation in natural scenes," *J. Vis.*, vol. 5, no. 8, p. 692, 2005.
[53] X. Sun, X. Yang, S. Wang, and M. Liu, "Content-aware rate control scheme for HEVC based on static and dynamic saliency detection," *Neurocomputing*, vol. 411, pp. 393–405, Oct. 2020.
[54] R. Yang, M. Xu, Z. Wang, Y. Duan, and X. Tao, "Saliency-guided complexity control for HEVC decoding," *IEEE Trans. Broadcast.*, vol. 64, no. 4, pp. 865–882, Dec. 2018.
[55] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for HEVC-MSP," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 155–170, Jan. 2018.
[56] M. Xu, Y. Liu, R. Hu, and F. He, "Find who to look at: Turning from action to saliency," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4529–4544, Sep. 2018.
[57] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[58] *How Long is Call of Duty: Black Ops II?* Accessed: Apr. 16, 2020. [Online]. Available: http://howlongtobeat.com/game.php?id=1472
[59] *The Most Popular Shooter Video Games Right Now*. Accessed: Apr. 16, 2020. [Online]. Available: https://www.ranker.com/list/most-popular-shooter-video-games-today/ranker-games
[60] *Call of Duty Franchise Game Sales Statistics*. Accessed: Apr. 16, 2020. [Online]. Available: http://www.statisticbrain.com/call-of-duty-franchise-game-sales-statistics/

[61] *Rayman Legends*. Accessed: Apr. 16, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Rayman_Legends
[62] S. Xie, Y. Xu, Y. Li, Q. Shen, Z. Ma, and W. Zhang, "Perceptually optimized quality adaptation of viewport-dependent omnidirectional video streaming," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 146–160, Jan. 2020.
[63] Y. Zhou, L. Tian, C. Zhu, X. Jin, and Y. Sun, "Video coding optimization for virtual reality 360-degree source," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 118–129, Jan. 2020.
[64] M. Dick, O. Wellnitz, and L. Wolf, "Analysis of factors affecting players' performance and perception in multiplayer games," in *Proc. 4th ACM SIGCOMM Workshop Netw. Syst. Support Games (NetGames)*, 2005, pp. 1–7.
[65] F. Pakdaman, M. A. Adelimanesh, M. Gabbouj, and M. R. Hashemi, "Complexity analysis of next-generation VVC encoding and decoding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3134–3138.
[66] S. Kuanar, K. R. Rao, and C. Conly, "Fast mode decision in hevc intra prediction, using region wise CNN feature classification," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2018, pp. 1–4.
[67] S. Kuanar, K. R. Rao, M. Bilas, and J. Bredow, "Adaptive CU mode selection in HEVC intra prediction: A deep learning approach," *Circuits, Syst., Signal Process.*, vol. 38, no. 11, pp. 5081–5102, Nov. 2019.
[68] E. Hosseini, F. Pakdaman, M. R. Hashemi, and M. Ghanbari, "Fine-grain complexity control of HEVC intra prediction in battery-powered video codecs," *J. Real-Time Image Process.*, pp. 1–16, Jul. 2020.

HAMED AHMADI received the master's degree in artificial intelligence from the K. N. Toosi University of Technology, in 2011, and the Ph.D. degree in information technology from the University of Tehran, in 2016. He is currently a Principal Machine Learning Researcher with Oracle Labs. His research interests include user behavior modeling, cloud gaming, and video streaming.

SAMAN ZADTOOTAGHAJ (Graduate Student Member, IEEE) received the bachelor's degree from IASBS and the master's degree in information technology from the University of Tehran. He is currently a Researcher with the Quality and Usability Laboratory, Technische Universität Berlin, working on modeling the gaming quality of experience under the supervision of Prof. Dr.-Ing. S. Möller. He worked as a Researcher with the Telekom Innovation Laboratories of Deutsche Telekom AG from 2016 to 2018 as part of European project, QoE-Net. His research interests include subjective and objective quality assessment of Computer-Generated content. He is also the Chair of the Computer-Generated Imagery (CGI) Group, Video Quality Expert Group (VQEG).

FARHAD PAKDAMAN received the B.Sc. degree in computer engineering from the University of Mazandaran, in 2011, and the M.Sc. and Ph.D. degrees in computer engineering from the University of Tehran, in 2013 and 2019, respectively. From 2014 to 2017, he has served as a Lecturer for the University of Mazandaran. Since January 2018, he has been a Researcher with the Signal Analysis and Machine Intelligence (SAMI) research group, Tampere University, Finland. He has been a Reviewer of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), IEEE ACCESS, and *Multimedia Tools and Applications*. His research interests include multimedia systems and applications, video processing and compression, high performance computing, power-aware computing, and application specific architectures.

MAHMOUD REZA HASHEMI (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Tehran, and the Ph.D. degree from the University of Ottawa, Canada. He is currently a tenured Associate Professor with the School of Electrical and Computer Engineering, University of Tehran. He is also the Co-Founder and the Director of the Multimedia Processing Laboratory (MPL). His research interests include multimedia systems and networking specifically cloud gaming and applied AI for multimedia systems. He was a recipient of the ICME Quality Reviewer awards in 2011 and 2013. He is also an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

SHERVIN SHIRMOHAMMADI (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Ottawa, Canada. He is currently a Professor with the School of Electrical Engineering and Computer Science, University of Ottawa. He is also the Director of the Distributed and Collaborative Virtual Environment Research Laboratory, doing research in applied AI for multimedia systems and networks, specifically video systems, gaming systems, and multimedia-assisted healthcare systems. The results of his research, funded by more than $14 million from public and private sectors, have led to over 350 publications, three best paper awards, over 70 researchers trained at the postdoctoral, Ph.D., and master's levels, holds over 20 patents and technology transfers to the private sector, and a number of awards. He is also a Fellow of the IEEE for contributions to multimedia systems and network measurements and a Lifetime Senior Member of the ACM. He was the winner of the 2019 George S. Glinski Award for Excellence in Research and the University of Ottawa Gold Medalist. He is also a licensed Professional Engineer in Ontario. He is also the Editor-in-Chief of the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, and an Associate Editor of *ACM Transactions on Multimedia Computing, Communications, and Applications*, having been numerously recognized as the Associate Editor of the year by both of these and other journals.