This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3118003, IEEE Internet of Things Journal

Deepview: Deep Learning based Users Field of View Selection in 360° Videos for Industrial Environments

#Muhammad Irfan, #Khan Muhammad, *Member, IEEE,* Muhammad Sajjad, Khalid Malik, *Senior Member, IEEE,* Faouzi Alaya Cheikh, Joel J. P. C. Rodrigues, *Fellow Member, IEEE,* Victor Hugo C. de Albuquerque, *Senior Member, IEEE*

***Abstract*—The industrial demands of immersive videos for virtual reality/augmented reality applications are crescendo, where the video stream provides a choice to the user viewing object of interest with the illusion of "being there". However, in industry 4.0, streaming of such huge-sized video over the network consumes a tremendous amount of bandwidth, where the users are only interested in specific regions of the immersive videos. Furthermore, for delivering full excitement videos and minimizing the bandwidth consumption, the automatic selection of the user's Region of Interest in a 360° video is very challenging because of subjectivity and difference in contentment. To tackle these challenges, we employ two efficient convolutional neural networks for salient object detection and memorability computation in a unified framework to find the most prominent portion of a 360° video. The proposed system is four-fold: preprocessing, intelligent visual interest predictor, final viewport selection, and virtual camera steerer. Firstly, an input 360° video frame is split into three Field of Views (FoVs), each with a viewing angle of 120°. Next, each FoV is passed to object detection and memorability prediction model for visual interestingness computation. Further, the FoV is supplied as a viewport, containing the most salient and memorable objects. Finally, a virtual camera steerer is designed using enriched salient features from YOLO and LSTM that are forwarded to the dense optical flow to follow the salient object inside the immersive video. Performance evaluation of the proposed system over our own collected data from various websites as well as on public datasets indicates the effectiveness for diverse categories of 360° videos and helps in the minimization of the bandwidth usage, making it suitable for industry 4.0 applications.**

***Index Terms*—Virtual reality, Industry 4.0, Deep learning, Saliency, Immersive videos, AR industry, View selection, IoT**

I. INTRODUCTION

A 360° camera provides a complete view of the immersive world, which makes it dominant over the traditional cameras. These cameras are adopted by the current emerging technologies and industrial applications in Virtual Reality (VR) [1, 2] and Augmented Reality (AR) [3, 4] because of their delightful experience to the users. Due to its high demand, 360° videos are also created and supported by the tech and social media giants, i.e., Facebook [5], YouTube, and Google [6]. Recently, more than one million of 360° video contents and 25 million 360 images are uploaded to Facebook. YouTube also started supporting and creating 360° content since 2015 that delivers exotic viewing experience to the end-users. Similarly, Google also provides development toolkits for multiple platforms to create 360° content that could be watched over smartphones. Further, in the industrial sector, the sales of VR and Mixed Reality headsets likewise boosted the popularity of 360° contents. In comparison to conventional videos, 360° videos give the users an exciting experience through the illusion of being there in the virtual environment [7, 8]. This new video genre has attracted users, huge industries, and researchers, however, at the same time created new challenges in its exploration for various applications [9]. In 360° contents, different problems are encountered due to the large field of view, bandwidth consumption [10], and the limited FoV of the human perception towards such visual contents. Foremost, it is very strenuous for a user to find the choice of "where to look" because of spacious coverage of a 360° video. Specific manual techniques make it possible to navigate current FoV in a 360° video. These involves mouse clicks and HMDs, where the sensors in such devices navigate in the video with the help of head movements. A typical example of manually viewing a 360° video using HMD is shown in Figure 1. But these methods produce mental stress and VR sickness where viewer feels discomfort while watching a 360° video [11]. To overcome these challenges, automatic virtual camera for 360° videos [12] is an appealing field where novel techniques are designed to

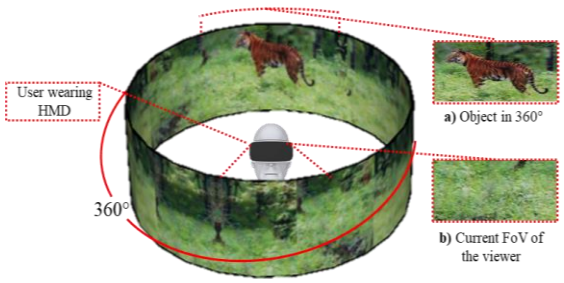

Fig. 1. Spherical view of 360° frame using Head-Mounted Display (HMD) device, where a) shows the object of interest in a 360 frame, while b) is the current FoV of the viewer.

Manuscript received February 10, 2021; Revised July 1, 2021; Accepted August 27, 2021; Published: XXXX. This research is supported by: (1) an ERCIM 'Alain Benoussan' Fellowship Programme under the Contract 2019–40 and (2) by Color and Visual Computing Lab, Department of Computer Science, NTNU, Gjøvik, Norway. The work of Joel J. P. C. Rodrigues was partially funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the Project UIDB/50008/2020; and by Brazilian National Council for Scientific and Technological Development - CNPq, via Grant No. 313036/2020-9. (*Corresponding authors: Khan Muhammad and Muhammad Sajjad*). (#Muhammad Irfan and #Khan Muhammad contributed as co-first authors)

Muhammad Irfan and Khan Muhammad are with Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Software, Sejong University, South Korea (e-mail: irfantahir301@gmail.com; khan.muhammad@ieee.org).

Muhammad Sajjad is with the Digital Image Processing Laboratory, Department of Computer Science, Islamia College Peshawar, Peshawar 25000, Pakistan and also with the Color and Visual Computing Lab, Department of Computer Science (IDI), Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway (e-mail: muhammad.sajjad@icp.edu.pk; muhammad.sajjad@ntnu.no).

Khalid Malik is with the Department of Computer Science & Engineering, Oakland University, Rochester, USA (e-mail: mahmood@akland.edu).

Faouzi Alaya Cheikh is with Color and Visual Computing Lab, Department of Computer Science (IDI), Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway (e-mail: faouzi.cheikh@ntnu.no).

Joel J. P. C. Rodrigues is with the Federal University of Piauí (UFPI), 64049-550 Teresina-PI, Brazil and also with Instituto de Telecomunicações, 6201-001 Covilhã, Portugal (e-mail: joeljr@ieee.org).

Victor Hugo C. de Albuquerque is with the Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza, 60811-905, Brazil (e-mail: victor.albuquerque@ieee.org).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3118003, IEEE Internet of Things Journal

minimize bandwidth usage and process an unedited video for generating visually alluring and pleasant events.

In the current literature, there exists several methods for virtual camera selection in 360° videos. Further, recent deep learning approaches such as reinforcement learning has also been explored for big data analytics [13-15]. Further, Yu-Chuan et al. [16] proposed a conventional algorithm that creates a virtual camera within 360° video for controlling the viewing angle of the viewer to watch 360° videos. However, their method lacks salient object detection in the video. Similar work is presented by Hou-Ning et al. [17] via leveraging deep learning-based approach called "deep 360 pilot", an agent that navigates viewing angle in 360° sports videos. However, their method is only adaptable for sport videos (skateboard). Moreover, for the wild 360° videos, a virtual viewing angle technique is developed by Cheng [12] via computing saliency-based heat maps for predicting the most salient scenes. Xu et al. [18] utilized eye-gaze data, measuring the saliency score of the object that helps to control "where to look" of the viewer in the 360° videos. The current problem of "where to look" is further studied by Li et al. [19] by proposing a virtual camera called "viewport". All these methods are domain-specific that only work for sports and wild videos where these systems have limitations dealing with other diverse 360° video categories. Furthermore, 360° videos are the primary source of entertainment and to the best of our knowledge, no such system exists that finds interesting and visually pleasant FoV in 360° videos.

In this paper, we propose "Deepview" a novel deep learning-based intelligent visual interest predictor (IVIP) architecture, which creates a virtual camera presenting visually interesting and pleasant scene in 360° videos. In the proposed system, an input 360° video is split into three 120° FoV's, where each scene is passed to the IVIP architecture for saliency computation. Deep learning based IVIP architecture calculates a score for each FoV based on the memorability and objects present in the scene. FoV with highest saliency score is further processed to compute the visual features using YOLO and LSTM. Next, dense optical flow controls the viewing angle of the virtual camera via tracking the salient objects inside an immersive video. For the evaluation of Deepview, we have included well-known 360° video categories, such as sports, entertainment, tour, wild, and cartoons. The proposed system also minimizes the bandwidth consumption supplying only a salient portion of the immersive contents. Following are the main contributions of this work.

1. A 360° video provides coverage of the entire surrounding environment that makes it difficult for a user to choose the choice of "where to look". It requires physical efforts and creates mental stress for users when the region of interest is outside of the current FoV. To tackle these challenges, we propose an intelligent and novel framework that finds the salient object and controls viewing angle by following the object inside the immersive videos.
2. 360° videos are main source of the entertainment, however, manually controlling viewing angles inside these contents is very tedious. Existing methods rely on hand-crafted heuristics, which produce VR sickness to the viewer. To overcome these challenges, two Convolutional Neural Networks (CNNs) are used to detect salient objects and compute their memorability score. Further, the measured scores of the salient objects are intelligently fused to find the most prominent FoV in the 360° videos.
3. To automatically control motion of the virtual camera inside a 360° video, we extended a state-of-the-art (SOTA) CNN model (i.e., ROLO). For the virtual camera steering, YOLO and LSTM are used to extract visual features and learn sequence pattern, respectively, providing the location of the objects. Finally, the location of the salient objects is passed to the dense optical flow to control the viewing angle of the virtual camera inside a 360° video.

The rest of the paper is organized in three sections. Section II consists of the proposed system, where the detailed overview and flow of the proposed framework is described. Section III describes the experimental results, where the system efficiency is evaluated using different experimental schemes comparing with other SOTA techniques. Section IV wraps up the paper with conclusion and future directions.

II. Proposed Framework

The proposed methodology is divided into four main steps: A) the mechanism of splitting a 360° video into 120° FoVs as a preprocessing, B) IVIP describes CNN architectures used to compute the saliency score of the FoVs, C) final viewport selection, and D) virtual camera steerer that allows to control viewing angle of the user based on the salient object motion. Figure 3 shows a detailed overview of the proposed framework.

A. Preprocessing

In computer vision, resolution is referred to the video dimensions but in the case of 360° videos, it is a bit complicated due to its panoramic view. In immersive videos, contents are stretched over 360° horizontally and 180° vertically, while the whole scene is stretched between two eyes of the viewer. This phenomenon enables the viewer to freely move inside 360° videos. However, due to the limited normal field of view (NFoV) of humans and VR devices, only 120° FoV can be seen by a viewer at the same time. As a result, the viewer misses most of the salient objects and entertainment events in the immersive videos. To overcome these challenges and provide high-resolution FoV to the viewer, an input 360° frame is split into three FoVs, each with a 120° view. Further, 360° videos are created in different resolutions ranging from 2k to 16k resolutions as shown in Figure 2. For the user's NFoV, the input

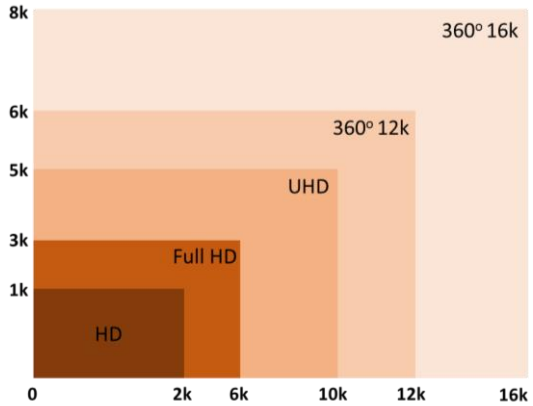

Fig. 2. Resolution of different standard displays and 360° videos.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3118003, IEEE Internet of Things Journal

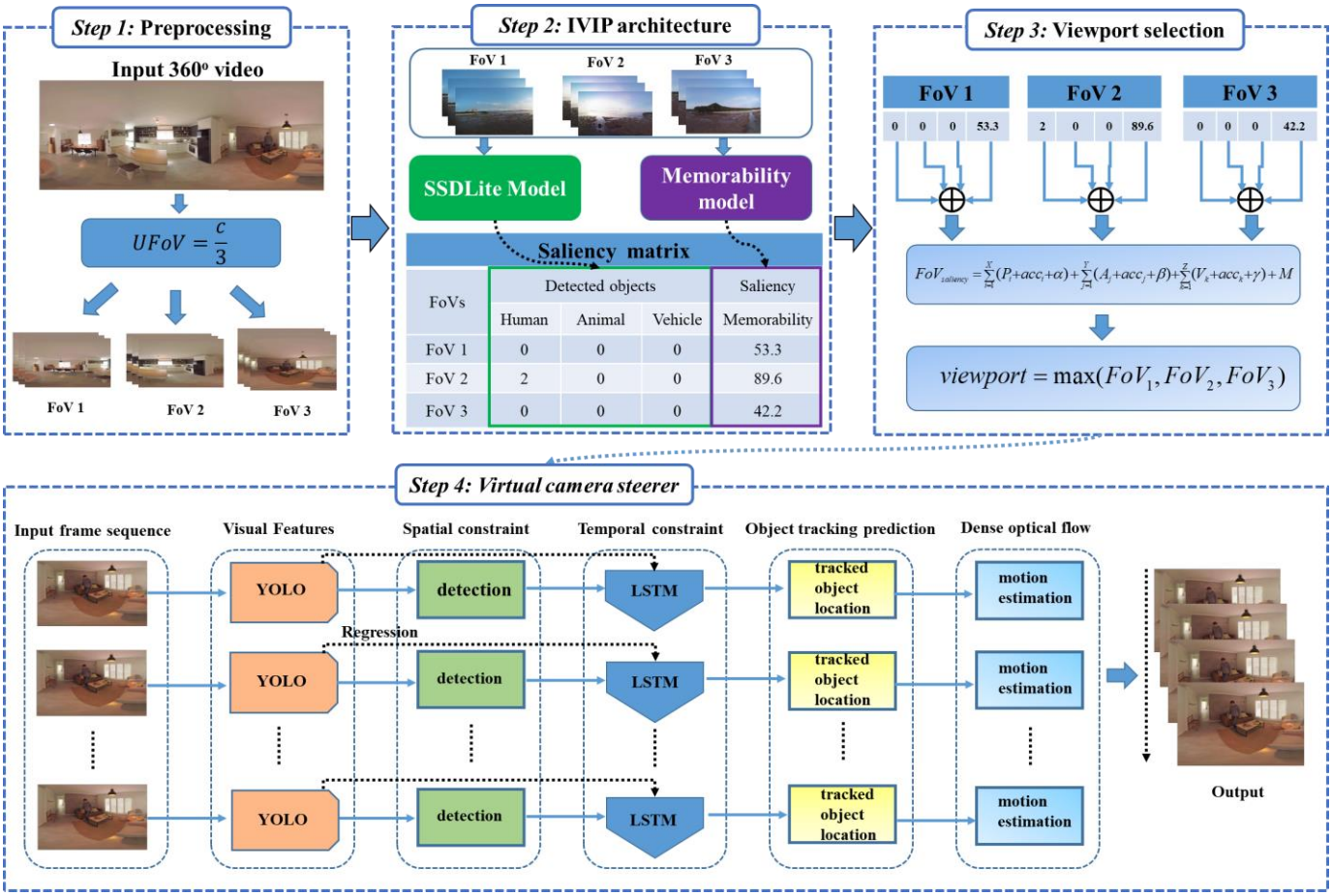

Fig. 3. Detailed overview of the proposed framework. *Step 1*: input 360° video frames are split into three 120° FoVs. *Step 2*: IVIP architecture predicts the saliency object based on the memorability score and objects present in each FoV. Objects and memorability score of each FoV are stored into saliency matrix. *Step 3*: Objects and memorability score of each FoV are fused and forwarded to a viewport selection module, where the FoV with high saliency is selected. *Step 4*: High-level features are extracted from the viewport using visual features of the YOLO and LSTM for objects motion estimation, where motion of the virtual camera is controlled using dense optical mechanism inside a 360° video.

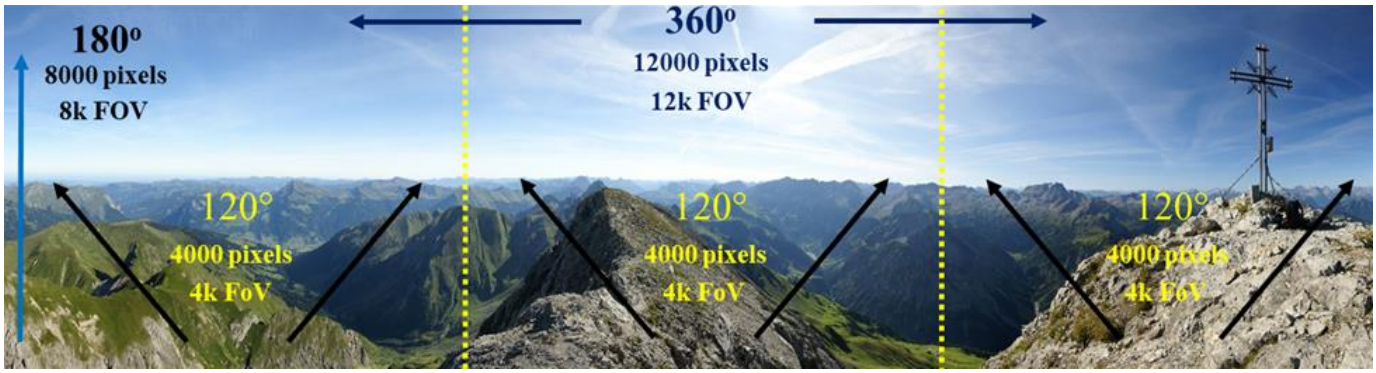

Fig. 4. Panoramic view of a 360° frame. Each input 360° frame is split into three FoVs while its resolution is adjusted according to the size of the input frame, e.g., the input 360° 12k frame is split into three 120 FoVs with 4k resolution.

frame is automatically adjusted according to the input resolution. However, the resolution of the user's NFoV varies as 360° videos come in various sizes. Table 1 shows different NFoV for various input sizes of the 360° videos where Figure 4 shows a panoramic view of the 360° frame.

B. Intelligent Visual Interest Predictor Architecture

In 360° videos, interesting FoV is a 120° view that is enriched with objects and memorable sceneries depending upon the video category [20]. The 360° videos are divided into several categories, including sports, entertainment, and tour videos, among others. A survey conducted by Nikon [21] suggests that 90% of the immersive consumers are interested in the 360° videos because of the improved experience. In the survey, it is also stated that 60% of the immersive consumers prefer to watch sports and travel contents in a 360° view as compared to the traditional videos. Moreover, the survey states that entertainment videos gained an interest of 55% as compared to other categories of immersive videos. The IVIP architecture is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3118003, IEEE Internet of Things Journal

further divided into two subsections i) Salient Object Detection using SSDLite and ii) Memorability Measurement.

i. Salient Object Detection using SSDLite

The common aspect of all these 360° videos in which the viewer is interested are salient objects and stunning sceneries. To select the FoV with salient objects, it is necessary to detect different types of objects in each scene. Towards this end, we employed an existing deep learning object detection model called SSDLite [22], which is a modified version of a Single Shot Detector (SSD) [23]. In contrast with SSD, SSDLite is lightweight version with modified kernel size including depth and pointwise separable convolutions that make the model efficient in terms of accuracy and time complexity. In the proposed system, to retain the object shape and prevent it from deformation, the input 360° video is converted into an equilateral form. Then, we employ a pre-trained SSDLite model where each FoV is fed, producing a vector consisting of salient objects classes and their corresponding confidence values. In the output vector, human has the top precedence followed by animals and vehicles.

TABLE 1
RESLUTION OF DIFFERENT 360° VIDEOS

Horizontal View	Vertical View	Resolution of NFoV of 120°
2000	1000	**~ 720x1000**
4000	2000	**~ 1320x2000**
6000	3000	**~ 2000x3000**
8000	4000	**~ 2600x4000**
10000	5000	**~ 3320x5000**
12000	6000	**~ 4000x6000**
16000	8000	**~ 5200x8000**

ii. Memorability Measurement

Most images have certain characteristics that attract human attention and are easier for them to remember as compared to other humans. These images contain notable events including people and objects with natural landscapes. In the literature, researchers have suggested many techniques that measure the memorability of images using different approaches [24]. A detailed study of these approaches suggests that images containing salient events or objects have popularity, aesthetics, emotion, and memorability. In the context of 360° videos, objects enriched scenes always fascinate viewers thus, memorability scores of these scenes are very high, enabling efficient detection of the salient scenes in immersive contents. In addition, we used an existing SOTA hybrid-AlexNet model that is fine-tuned with a large annotated image memorability dataset called LaMem [25]. This model can recognize memorability score of various classes including humans, animals, and beautiful natural sceneries. In the proposed framework, we compute the memorability score of each FoV based on the trained weights and fused it with other modules for salient and interesting FoV selection. Finally, objects with their accuracy and memorability score are mapped into the saliency matrix for each FoV using Eq. 1.

$$FoV_{saliency} = \sum_{i=1}^{X}(\phi_i + acc_i + \alpha) + \sum_{j=1}^{Y}(\varphi_j + acc_j + \beta) + \sum_{k=1}^{Z}(\delta_k + acc_k + \gamma) + M \quad (1)$$

Here ϕ_i denotes the i^{th} person, acc_i is confidence value of ϕ_i, φ_j denotes j^{th} animal, acc_j is confidence value of φ_j, δ_k denotes k^{th} vehicle, and acc_k is confidence value of δ_k. Further, X, Y, Z, and M is the total number of persons, animals, vehicles, and memorability score of the FoV, respectively. Moreover, α, β, and γ are the balancing weights assigned to three classes: persons, animals, and vehicles. Description of parameters are given in Table 2. Further, from the study conducted by Rai et al. [26], it can be perceived from the heatmaps that users were more interested in humans as compared to other categories. Therefore, we introduced a balancing weight 1 where half of the weight is assigned to α, and the remaining half is distributed at a ratio of 0.3 and 0.2 between β and γ, respectively. Moreover, to reduce false detection of the SSDLite model, the confidence value *acc* is limited to a fixed value. All the objects with lower *acc* are ignored. This mechanism helps us to improve the system's effectiveness in terms of accuracy and false detection.

TABLE 2
DESCRIPTION OF PARAMETERS

Parameter	Description
A	Accuracy
P	Precision score
R	Recall
F1	F1-score
M	Memorability score
Aes	Aesthetic score
Mem	Memorability score
Obj	Salient object
Mem+Aes	Fusion of memorability and aesthetic score
Aes+Obj	Fusion of aesthetic salient object score
Mem+Obj	Fusion of salient object and memorability score
acc	Confidence score
ϕ	Person
φ	Animal
δ	Vehicle
α, β, and γ	Balancing weights

C. Viewport Selection

The most important phase of our proposed system is the pleasant and exciting FoV selection based on the saliency matrix generated via IVIP architecture. Eq. 2 chooses the most important FoV as a viewport for the user.

$$viewport = max(FoV_1, FoV_2, FoV_3) \quad (2)$$

Here FoV_1, FoV_2, and FoV_3 are saliency score of each 120° FoV. The equation provides saliency score of the salient objects present in each FoV. The *Viewport* is the maximum saliency score based FoV that is used as a final NFoV for the users. The highest saliency score-based viewport is further passed to the virtual camera steerer for controlling the viewing angle of the user via following the salient object inside the 360° video.

D. Virtual Camera Steerer

After the measurement of most salient FoV, our next aim is to create a virtual camera that could steer the salient FoV for the viewer inside an immersive video. Inspired by the recent success of the recurrent neural networks in various domains of computer vision, we extend a SOTA deep model ROLO [27] to control the virtual camera inside immersive videos. For the motion estimation of the salient FoV, we obtain the rich visual features and primary location of the objects via YOLO [81]. At the end, YOLO adopts fully connected layers to transform regressing features representation into regions predictions. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3118003, IEEE Internet of Things Journal

returned tensor is coded as $U \times U \times (B \times 5 \times C)$ where $U \times U$ indicates the number of image slices, B is the number of bounding boxes in each slice with 5 location including row, column, width, height, and confidence score C. C represents the class label of each bounding box. In the proposed work, we adopted the same setup as in the YOLO model and set $U = 7$, $B = 2$, and $C = 20$. Our focus is to detect the salient object and to control the viewing angle by following the object. Therefore, we drop the class labels and confidence score as:

$$B_t = (0, x, y, w, h, 0) \quad (3)$$

Where x, y, w, and h represent the x-axis, y-axis, width, and height of the bounding box, respectively.

In the last, we add LSTM with two inputs namely, features from the fully connected layers and detection from $B_{t,i}$. At a given time-step t, feature vector of length 4096 is extracted referred as X_t. Another input to the LSTM is the last time-step S_{t-1}. Output of the LSTM provides locations of the object, which are passed from the dense optical flow to estimate the overall motion of the objects in the FoV. The dense optical flow algorithm works very efficiently to estimate the motion of interesting features via comparison of the two adjacent frames in the video. Our final goal with the dense optical flow is to find the x and y coordinates of the virtual camera following the salient objects. The $m \times n$ window (i.e., size of the salient FoV) is taken assuming that all the pixels of the window have the same motion as represented in Eq. 4.

$$\sum_{i=1}^{m \times n} I_x(q_i)\, V_x + I_y(q_i)\, V_y = -\, I_t(q_i) \quad (4)$$

Here $m \times n$ is the total number of pixels inside the window and $I_x(q_i)$, $I_y(q_i)$, and $I_t(q_i)$ are the partial derivatives of the frame I with respect to position (x, y) and time interval t for pixel q_i at the current time. Hence our final goal is to control the motion of the overall virtual camera inside an immersive video via following the salient features. Suppose V_x and V_y are the vertical and horizontal axis of the virtual camera such that:

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix} = \begin{bmatrix} \sum_i I_x(q_i)^2 & \sum_i I_x(q_i) I_y(q_i) \\ \sum_i I_y(q_i) I_x(q_i) & \sum_i I_y(q_i)^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i I_x(q_i) I_t(q_i) \\ -\sum_i I_y(q_i) I_t(q_i) \end{bmatrix} \quad (5)$$

where $V_x = d_x/d_t$ represents the movement of the virtual camera along x-axis with time t and $V_y = d_y/d_t$ denotes the movement of the virtual camera along y-axis over time t. Further, in the proposed system, we set the virtual camera height $h = 640$ and width $w = 480$ for all set of experiments.

III. Experimental Results

This section describes the experimental setup and evaluation of the proposed system on different sets of videos. In the experimental evaluation, we conducted various tests for both the salient object detection module and virtual camera steerer module to find the adaptability of the proposed system in various domains of immersive contents. The detailed discussion of the collected videos and the results obtained from the proposed system are presented in the following subsections.

A. Experimental Setup and Datasets

The proposed system is implemented using Python version 3.6 and an open-source image processing library OpenCV version 4.0. Other necessary libraries that were used for pre-processing, training, and visualization include Numpy, Keras, Tensorflow (GPU version), Caffe (compiled for Python), Matplotlib, Scikit-image, and Scikit-learn. Several videos that belong to different categories including sports, tour, entertainment, cartoon, and documentary were downloaded from YouTube to evaluate the performance of the proposed system. These videos contain both static viewpoint (SVP) and moving viewpoint (MVP) where in SVP videos, the salient objects are static and in MVP videos, objects move around in the 360°. All videos were downloaded in equirectangular format with .mp4 extension at 30 frames per second (fps). The detail about each video is given in Table 3. To further evaluate the performance of the proposed system, we compared it with other SOTA systems using Salient-360 [26] dataset and PVS-HM [28] dataset. The Salient-360 dataset consists of 19 videos containing eye tracking and head movements of 57 participants (32 males and 25 females). Later, both features (features from eye tracking and head movements) of participants watched FoV are fused to generate saliency maps. On the other hand, the PVS-HM dataset only consists of eye-tracking data of 58 participants.

B. Performance Evaluation over Salient Object Detection

To analyze the performance of the proposed system, we carried out different schemes of experiments combining various deep learning models including memorability (Mem), SSDLite (Obj), and aesthetic (Aes). The evaluation metrics consist of accuracy A, precision P, recall R, and harmonic mean of the precision and sensitivity (F). Firstly, the aesthetic score of all three FoVs are calculated, and the FoV with a high aesthetic score is presented to the viewer. However, due to the variation in contents, objects, scenes, and lighting conditions in the videos, the best A score is only 0.52. Similarly, in the next test, the FoV carrying highest score is presented using memorability model. The memorability model produces the highest score for A as shown in Table 4, which is 0.56, but the FoV is not convincing in real scenario as shown in Figure 6. In the proposed system, we also utilized SSDLite object detection model presenting FoV based on the salient objects. The model raised overall A, P, R, and F scores to 0.60, 0.58, 0.55 and 0.61, respectively. Among these deep learning models, this was the highest possible score on a single network.

TABLE 3
VIDEOS USED IN THE EVALUATION

Video Title	Type	Focus Point	Starting offset	FoV
360° Degree Kitchen Home Tour	SVP	Persons	0:01	4k
Kitchen 360° test tour	MVP	Person	0:01	4k
Learning to skateboard in Venice (in 360° video)	MVP	Persons	0:01	2k
GoPro VR: Tahiti Surf with Anthony Walsh and Matahi Drollet	MVP	Persons	0:15	4k
Lions 360° National Geographic	MVP	Animal	0:07	2k
Clash of Clans 360°: Experience a Virtual Reality Raid	SVP	Cartoon	0:04	2k
360° Underwater National Park National Geographic	MVP	Animal	0:04	2k

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3118003, IEEE Internet of Things Journal

Fig. 5. Sample frames from the videos, where the green squares show the correct and the red squares show the incorrect FoV predicted by the proposed system.

To further improve the performance of the proposed system, we fused these networks in different orders. The highest possible *A*, *P*, *R*, and *F* scores are 0.72, 0.70, 0.68, and 0.71, obtained via the fusion of memorability and SSDLite model. Representative frames of the salient objects using the proposed system are shown in Figure 5. The overall performance of the single and fusion of different networks is illustrated in Figure 6.

C. Performance Evaluation of the Virtual Camera Steerer

Immersive videos are dominant over the traditional videos that provide user the choice to move freely inside 360° environments, enabling them to be the primary source of entertainment in VR. To provide the users full excitement inside the VR environment, the proposed method designed a virtual camera to follow the salient objects inside the 360° videos. For the evaluation of the proposed virtual camera steerer, we used seven videos (Section III-A) to validate the performance of the proposed system. The heat maps of the salient objects and the viewing angle along the current time is demonstrated in Figure 7. Further, we used a second set of experiments and performance indicators including true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), and accuracy to find the adaptability of the proposed system. We also compared the tracking capabilities of the proposed virtual camera steerer with SOTA trackers including Multiple Instance Learning (MIL), Kernelized Correlation Filters (KCF), Tracking Learning and Detection (TLD), MedianFlow, Mosse, CSRT, and GoTurn. Among these trackers, the worst performance is observed for the MedianFlow due to the tracking of the object's failure.

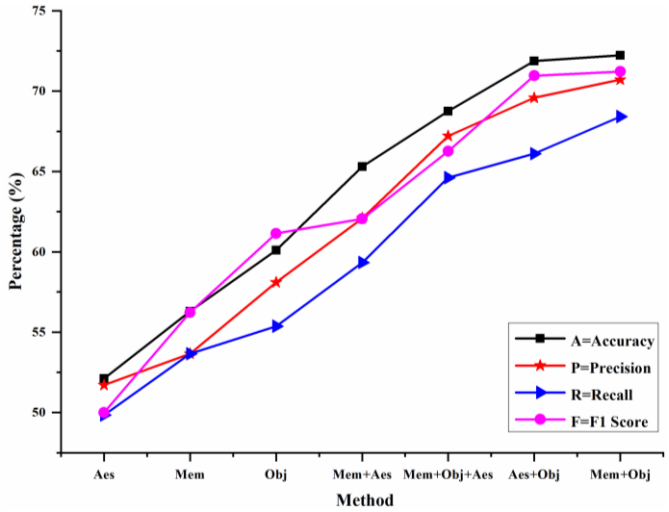

Fig. 6. Performance of the proposed system using single and fusion of different approaches.

On the other hand, GoTurn tracker is unable to handle occlusion problems in the various videos. Performance of the virtual camera steerer with other SOTA trackers is demonstrated in Table 4.

TABLE 4
EVALUATION OF VIRTUAL CAMERA STEERER WITH OTHER SOTA TRACKERS

Method	TPR (%)	TNR (%)	FNR (%)	Accuracy (%)
MIL	36.87	35.74	36.57	61.21
KCF	36.67	35.78	34.91	67.53
TLD	35.84	36.23	34.58	69.56
MedianFlow	39.21	40.75	37.57	61.21
Mosse	31.24	29.46	27.68	73.56
CSRT	19.27	21.35	20.78	81.24
GoTurn	38.47	39.67	38.67	62.37
Virtual Camera Steerer	**3.67**	**3.89**	**4.02**	**96.23**

D. Subjective Evaluation

We also conduct a user study (subjective evaluation) to compare and investigate the usefulness of the proposed system using the user interaction with 360° videos in more detail. The devices used in this study are Samsung S6-edge smartphone for playing the 360° videos and Samsung Gear VR HMD for presenting the 360° videos to the users. After watching the 360° videos, the final generated FoV is also evaluated by the participants. A total of 20 participants are recruited from different departments. The ages of these participants are between 20 to 40 years. In this scheme of experiments, first users are allowed to watch videos on HMD device and find the interesting FoV through head movement using manual methods (head movements). In the next step, the output video of the proposed system is played to the users where they watch a video without manually searching the FoV. Before presenting each video to the users, interesting FoV of all the videos are generated using the proposed system. After watching the videos, a questionnaire is given to the user to rate them based on their satisfaction. In the questionnaire, five different types of questions are given to validate the performance of viewing angle as: 1) Excellent, 2) Good, 3) Satisfactory, 4) Needs improvement, and 5) Poor. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3118003, IEEE Internet of Things Journal

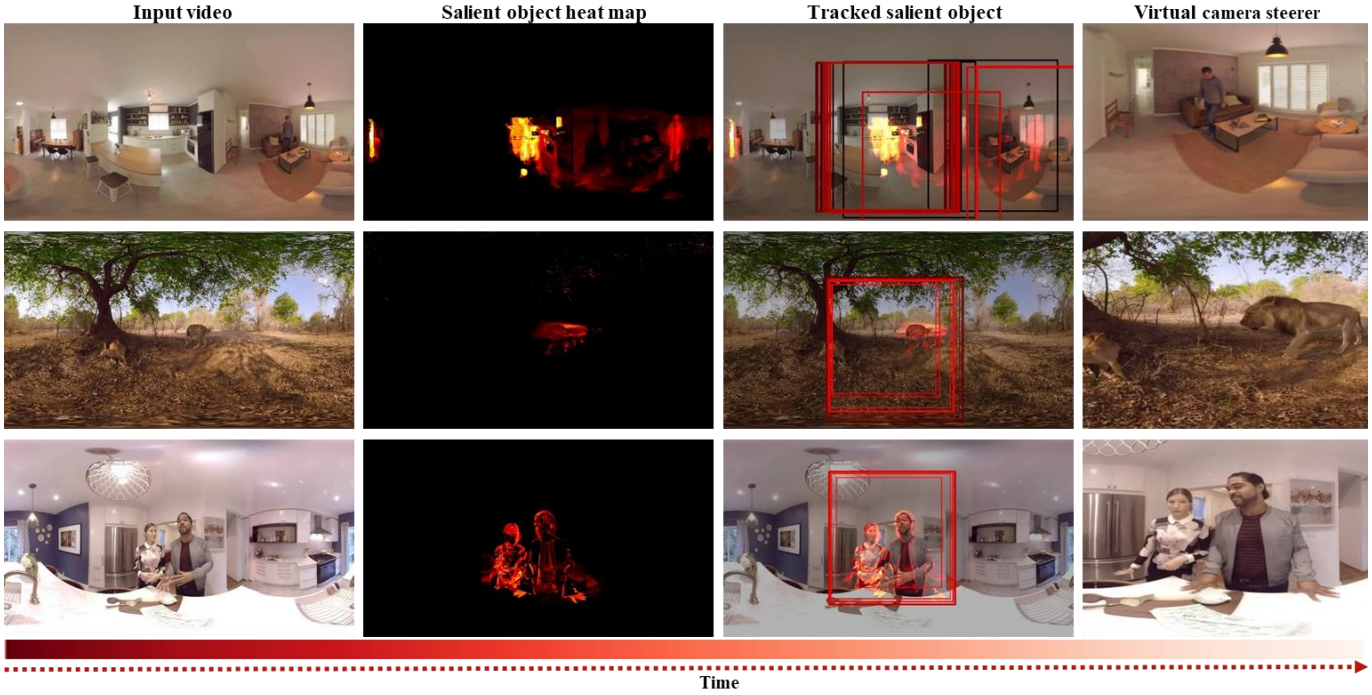

Fig. 7. Visual results of the proposed IVIP, where first column shows the input videos, second column shows the motion of the salient objects, third column shows the tracked salient objects, and fourth column shows the final virtual camera steerer view for the users.

TABLE 5
PERFORMANCE COMPARISION WITH STAT-OF-THE-ART METHODS

Videos	[17]				[12]				[16]				Deepview (Ours)			
	A	***P***	***R***	***F***	***A***	***P***	***R***	***F***	***A***	***P***	***R***	***F***	***A***	***P***	***R***	***F***
Video 1	0.55	0.54	0.49	0.51	0.53	0.54	0.53	0.54	0.52	0.53	0.51	0.53	0.75	0.73	0.66	0.73
Video 2	0.34	0.36	0.51	0.32	0.35	0.37	0.36	0.35	0.37	0.34	0.35	0.34	0.71	0.72	0.70	0.69
Video 3	0.84	0.87	0.79	0.82	0.83	0.84	0.79	0.80	0.38	0.36	0.34	0.33	0.69	0.71	0.67	0.72
Video 4	0.86	0.84	0.81	0.86	0.85	0.86	0.82	0.79	0.36	0.37	0.37	0.37	0.72	0.70	0.70	0.70
Video 5	0.35	0.34	0.30	0.31	0.37	0.33	0.33	0.35	0.86	0.87	0.85	0.83	0.71	0.69	0.71	0.69
Video 6	0.51	0.53	0.49	0.50	0.51	0.52	0.50	0.49	0.51	0.57	0.54	0.57	0.76	0.72	0.68	0.75
Video 7	0.53	0.36	0.33	0.34	0.34	0.34	0.34	0.31	0.35	0.39	0.37	0.34	0.72	0.70	0.69	0.70
Average	0.56	0.54	0.53	0.52	0.54	0.52	0.52	0.51	0.52	0.49	0.47	0.47	**0.72**	**0.71**	**0.68**	**0.71**

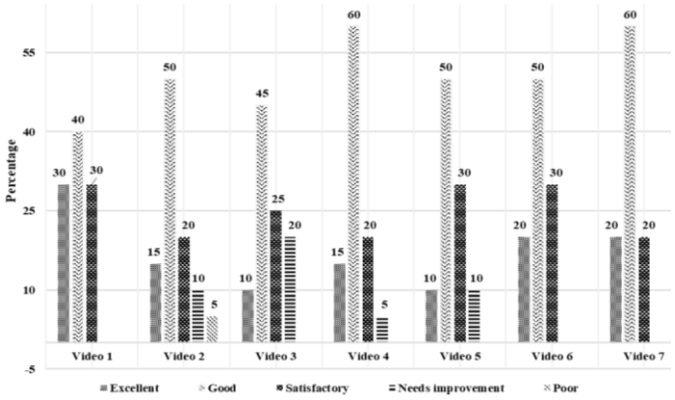

Fig. 8. Percentage of user's satisfaction after watching each video using HMD device.

participants then rate the video among the five options depending upon the user's comfort and satisfaction after experiencing the video. In both MVP and SVP, the proposed system attracts more users than the manually searching for salient object in the video using HMD device. The overall percentage of SVP videos are slightly higher than the MVP videos as the FoV in the SVP was constant and more stable than the MVP videos. The percentage of user's comfort and satisfaction on each video computed via the proposed system is illustrated in Figure 8.

E. Comparison with State-of-the-art Techniques

In this section, we validated the combined effect (salient object detection and controlling viewing angle inside 360° videos) with other SOTA methods including [12], [16], and [17]. Comparison of the proposed system with these SOTA methods [12][16][17] can be perceived from Table 5. Results of all the seven videos have been computed on SOTA methods. For video 1, the performance of [17] is better than those of [12] and [16], however the proposed system outperforms all three methods. In Table 5, method [12] clearly outperforms all other methods, nevertheless, the proposed method outperformed [17] and [16]. The worst time complexity is observed for the method [17] where the system is heavily dependent on the Faster R-CNN and the very high resolution of the 360° videos result in poor

TABLE 6
COMPARISION WITH OTHER STATE-OF-THE-ART TECHNIQUES

Method	Average time complexity	fps
[17]	1.24	~ 0.76
[12]	0.12	~ 8.04
[16]	0.14	~ 7.02
Deepview (Ours)	0.31	~ 3.07

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3118003, IEEE Internet of Things Journal

performance. Furthermore, the minimum time complexity is observed for the method [12] as its system is focusing on salient object in the 360° videos. From Table 6, it can be perceived that the proposed system has settled good tradeoff among existing systems on both time complexity and frames per seconds and at the same time, achieving best performance.

For the performance comparison of the proposed system with SOTA, we also used other performance evaluation metrics, including AUC-JuDD, correlation coefficient (CC), normalized scan-path saliency (NSS), and Kullback-Leiber (KL) divergence. The AUC-JuDD is the predicted saliency of the system and ground truth, where the score for a perfect match (ground truth and system predicted score) is 1. The CC computes the linear correlation of the predicted saliency and ground truth. The highest score for the CC is 1 that indicates an exact match of the predicted saliency and ground truth. Further, the NSS score deals with the average normalized accuracy of the saliency map, and a higher NSS score means more effectiveness of the system. KL score measures the total distribution of the predicted saliency over the ground truth, where 0 represents a perfect match between both scores. However, for the effective evaluation of the proposed system on two different datasets, we computed average AUC, CC, NSS, and KL score of both the datasets and compared with SOTA methods as shown in Table 7.

F. Quality of Service (QoS)

State-of-the-art methods focus on bandwidth minimization and latency rate of the 360° videos. They use various techniques for users' FoV to minimize bandwidth consumption. To evaluate this aspect of the proposed system (QoS performance) with respect to other SOTA methods, we created a simple server-client application using Python flask framework. In this set of experiments, we evaluated the latency and bandwidth consumption of the proposed method with other SOTA methods using local LAN and Wi-Fi services. The overall performance of our system with existing SOTA methods is illustrated in Table 8. In Table 8, the maximum bandwidth is consumed by [29], where the approach relies on the head and eye-tracking data of the users. On the other hand [30] consumed an average bandwidth of 13.01Mbps with an average latency rate of 10 ms. Our proposed system sends only salient and stunning FoV to the client, thereby consuming minimum bandwidth and latency rate among all the SOTA methods.

G. Limitations

The proposed system outperformed SOTA methods but it has certain limitations and is unable to recover from a number of failures. Firstly, IVIP only measures the saliency score at the start.

TABLE 7
PERFORMANCE EVALUATION WITH SOTA METHODS

Method	**Dataset**	**AUC**	**CC**	**NSS**	**KL**
[28]	PVS-HM	0.118	0.091	1.066	-
[31]	Salient-360	0.882	0.659	1.582	0.803
[32]	Salient-360	0.700	0.054	0.910	0.992
[33]	Salient-360	0.714	0.541	1.014	0.882
[34]	Salient-360	0.700	0.450	0.810	0.790
Deepview (Ours)	**PVS-HM/ Salient-360**	**0.913**	**0.725**	**1.627**	**0.677**

TABLE 8
AVERAGE BANDWIDTH AND LATENCY

Method	Average bandwidth (Mbps)	Average latency (ms)
CUR [29]	24.91	3
DR [29]	23.97	10
[29]	18.54	40
[30]	13.01	10
Deepview (Ours)	**12.45**	**12**

However, if there is a scene change in the video, our system is unable to detect the change and hence it cannot measure the saliency score of the new scene. Furthermore, virtual camera is based on dense optical flow that limits the performance when there is an abrupt change in the motion and occlusion of the objects. A 360° video provides vast field of view where the object moves freely. However, when the objects move to the edges of the virtual camera, the shapes are deformed, and the virtual camera is unable to detect the objects and control the viewing angle.

IV. CONCLUSION

This paper focused on the visual saliency prediction of the 360° videos to improve the user experience and minimize the bandwidth consumption while watching immersive videos. The input 360° frames were split into three 120° FOVs to predict salient objects and appealing scenes, and to reduce the overall bandwidth consumption over the network. We employed two CNN networks to extract saliency maps for this purpose from each FoV. Among the three FoVs, the most salient FoV was displayed to the viewer. Further, we extended a state-of-the-art CNN model by extracting visual features using YOLO and LSTM of salient objects. Moreover, the extracted features were passed to the dense optical flow algorithm to control the viewing angle of the user by following the salient object inside the 360° videos. We evaluated the performance of the proposed system on our own collected videos as well as on publicly available datasets. The extensive sets of experiments demonstrated the effectiveness of the proposed system in predicting more salient objects and improving QoS of the immersive videos, making it suitable for industry 4.0 applications. In the future, we have intention to minimize the number of CNN models to reduce time and computation complexity. Furthermore, instead of only following the salient objects inside the immersive contents, we will facilitate users to experience immerse events. Moreover, we are keen to increase the admissible number of frames that could deliver real-time experience to users.

REFERENCES

1. Lv, Z., et al., *BIM Big Data Storage in WebVRGIS.* IEEE Transactions on Industrial Informatics, 2020. 16(4): p. 2566-2573.
2. Lv, Z., et al., *Virtual Reality Smart City Based on WebVRGIS.* IEEE Internet of Things Journal, 2016. 3(6): p. 1015-1024.
3. Li, Z., Y. Wang, and X. Ji, *Monocular viewpoints estimation for generic objects in the wild.* IEEE Access, 2019. **7**: p. 94321-94331.
4. Livatino, S., F. Banno, and G. Muscato, *3-D Integration of Robot Vision and Laser Data With Semiautomatic Calibration in Augmented Reality Stereoscopic Visual Interface.* IEEE Transactions on Industrial Informatics, 2012. 8(1): p. 69-77.
5. Kopf, J., *360 video stabilization.* ACM Transactions on Graphics (TOG), 2016. 35(6): p. 195.
6. Anderson, R., et al., *Jump: virtual reality video.* ACM Transactions on Graphics (TOG), 2016. 35(6): p. 198.

7. Wang, D., K. Ohnishi, and W. Xu, *Multimodal Haptic Display for Virtual Reality: A Survey.* IEEE Transactions on Industrial Electronics, 2020. 67(1): p. 610-623.
8. Zhang, Y., et al. *Improving Quality of Experience by Adaptive Video Streaming with Super-Resolution*. in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*. 2020.
9. Caserman, P., A. Garcia-Agundez, and S. Goebel, *A Survey of Full-Body Motion Reconstruction in Immersive Virtual Reality Applications.* IEEE Transactions on Visualization and Computer Graphics, 2019: p. 1-1.
10. Zhang, X., et al., *Cooperative Tile-Based 360° Panoramic Streaming in Heterogeneous Networks Using Scalable Video Coding.* IEEE Transactions on Circuits and Systems for Video Technology, 2020. 30(1): p. 217-231.
11. Kim, H.G., et al. *Measurement of exceptional motion in VR video contents for VR sickness assessment using deep convolutional autoencoder*. in *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. 2017. ACM.
12. Cheng, H.-T., et al. *Cube padding for weakly-supervised saliency prediction in 360 videos*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
13. Xu, C., et al., *Renewable energy-aware big data analytics in geo-distributed data centers with reinforcement learning.* 2018. **7**(1): p. 205-215.
14. Lu, H., et al., *Edge QoE: Computation offloading with deep reinforcement learning for Internet of Things.* 2020. **7**(10): p. 9255-9265.
15. Wang, Y., et al., *Traffic and computation co-offloading with reinforcement learning in fog computing for industrial applications.* 2018. 15(2): p. 976-986.
16. Su, Y.-C. and K. Grauman. *Making 360 video watchable in 2d: Learning videography for click free viewing*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. IEEE.
17. Hu, H.-N., et al. *Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. IEEE.
18. Xu, Y., et al. *Gaze prediction in dynamic 360 immersive videos.* in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
19. Li, C., et al. *Viewport Proposal CNN for 360deg Video Quality Assessment.* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
20. Duan, L., et al. *Visual saliency detection by spatially weighted dissimilarity*. in *CVPR 2011*. 2011.
21. MELVILLE, N.Y., *Life in 360*. https://www.prnewswire.com/news-releases/study-reveals-that-majority-of-americans-are-ready-to-experience-life-in-360-300375057.html, 2016. Accessed on 16-01-2019.
22. Sandler, M., et al. *Mobilenetv2: Inverted residuals and linear bottlenecks*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
23. Liu, W., et al. *Ssd: Single shot multibox detector*. in *European conference on computer vision*. 2016. Springer.
24. Wang, L., et al. *Learning to detect salient objects with image-level supervision*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.
25. Zhang, X., et al. *Progressive attention guided recurrent network for salient object detection*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
26. Rai, Y., J. Gutiérrez, and P. Le Callet. *A dataset of head and eye movements for 360 degree images*. in *Proceedings of the 8th ACM on Multimedia Systems Conference*. 2017.
27. Ning, G., et al. *Spatially supervised recurrent convolutional neural networks for visual object tracking*. in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2017. IEEE.
28. Xu, M., et al., *Predicting Head Movement in Panoramic Video: A Deep Reinforcement Learning Approach.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. 41(11): p. 2693-2708.
29. Fan, C., et al., *Optimizing Fixation Prediction Using Recurrent Neural Networks for 360° Video Streaming in Head-Mounted Virtual Reality.* IEEE Transactions on Multimedia, 2020. 22(3): p. 744-759.
30. Zhang, Y., et al., *EPASS360: QoE-aware 360-degree Video Streaming over Mobile Devices.* IEEE Transactions on Mobile Computing, 2020: p. 1-1.
31. Zhu, Y., et al., *The Prediction of Saliency Map for Head and Eye Movements in 360 Degree Images.* IEEE Transactions on Multimedia, 2019: p. 1-1.
32. Cornia, M., et al. *A deep multi-level network for saliency prediction*. in *23rd International Conference on Pattern Recognition*. IEEE.
33. Jiang, M., et al. *Salicon: Saliency in context*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015. IEEE.
34. Lebreton, P. and A.J.S.P.I.C. Raake, *GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images.* Signal Processing: Image Communication, 2018. 69: p. 69-78.

Muhammad Irfan received his MS degree in Computer Science from Sejong University, Seoul, South Korea. He is working as a researcher in Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab). His research interest includes image and video processing, medical image analysis, intelligent transportation systems, computer vision, machine learning, and deep learning.

Khan Muhammad (Member, IEEE) received the Ph.D. degree in digital contents from Sejong University, Republic of Korea, in February 2019. He has been working as an Assistant Professor with the Department of Software since March 2019. He is currently the Director of the Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Sejong University, Seoul, South Korea. His research interests include intelligent video surveillance (fire/smoke scene analysis, transportation systems, and disaster management), medical image analysis, (brain MRI, diagnostic hysteroscopy, and wireless capsule endoscopy), information security (steganography, encryption, watermarking, and image hashing), video summarization, multimedia data analysis, computer vision, the IoT/IoMT, and smart cities.

Muhammad Sajjad received the master's degree from the Department of Computer Science, College of Signals, National University of Sciences and Technology, Rawalpindi, Pakistan in 2012, and the Ph.D. degree in digital contents from Sejong University, Seoul, South Korea in 2015. He is currently working as an ERCIM Research Fellow at NTNU, Norway. He is an Associate Professor with the Department of Computer Science, Islamia College University Peshawar, Pakistan. He is also the Head of the Digital Image Processing Laboratory with Islamia College University Peshawar. His primary research interests include computer vision, image understanding, pattern recognition, robotic vision, and multimedia applications. He has published more than 65 papers in peer-reviewed international journals and conferences. He is serving as a professional reviewer for various well-reputed journals and conferences.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3118003, IEEE Internet of Things Journal

Khalid Mahmood Malik (Senior Member, IEEE) is currently an Associate Professor with the School of Engineering and Computer Science, Oakland University, Rochester, MI, USA. His research interests include video and audio forensics, development of intelligent decision support systems using analysis of medical imaging and clinical text, secure multicast protocols for intelligent transportation systems, and automated ontology and knowledge graph generation. His research is supported by the National Science Foundation (NSF), Brain Aneurysm Foundation, and Oakland University.

Faouzi Alaya Cheikh received a BSc in electronics from l'Ecole Nationale d'Ingenieurs de Tunis in 1992, an MSc in Signal Processing in 1997 and a Dr. Tech. degree in Information Technology from Tampere Univ. of Technology (TUT), Tampere, Finland in 2004. Currently, he is working as full Professor at Department of Computer Science and Media Technology at Gjøvik University College in Norway. His research interests include e-Learning, machine learning, 3D imaging, image and video processing and analysis, video-guided intervention, biometrics, pattern recognition and content-based image retrieval.

Joel J. P. C. Rodrigues [S'01, M'06, SM'06, F'20] is a professor at the Federal University of Piauí, Brazil; and senior researcher at the Instituto de Telecomunicações, Portugal. Prof. Rodrigues is the leader of the Next Generation Networks and Applications (NetGNA) research group (CNPq), an IEEE Distinguished Lecturer, Member Representative of the IEEE Communications Society on the IEEE Biometrics Council, and the President of the scientific council at ParkUrbis – Covilhã Science and Technology Park. He was Director for Conference Development - IEEE ComSoc Board of Governors, Technical Activities Committee Chair of the IEEE ComSoc Latin America Region Board, a Past-Chair of the IEEE ComSoc Technical Committee (TC) on eHealth and the TC on Communications Software, a Steering Committee member of the IEEE Life Sciences Technical Community and Publications co-Chair. He is the editor-in-chief of the International Journal of E-Health and Medical Communications and editorial board member of several high-reputed journals (mainly, from IEEE). He has been general chair and TPC Chair of many international conferences, including IEEE ICC, IEEE GLOBECOM, IEEE HEALTHCOM, and IEEE LatinCom. He has authored or coauthored about 1000 papers in refereed international journals and conferences, 3 books, 2 patents, and 1 ITU-T Recommendation. He had been awarded several Outstanding Leadership and Outstanding Service Awards by IEEE Communications Society and several best papers awards. Prof. Rodrigues is a member of the Internet Society, a senior member ACM, and Fellow of IEEE.

Victor Hugo C. de Albuquerque [M'17, SM'19] is Professor and senior researcher at the Department of Teleinformatics Engineering / Graduate Program on Teleinformatics Engineering at the Federal University of Ceará, Brazil.
He has a Ph.D in Mechanical Engineering from the Federal University of Paraíba (UFPB, 2010), an MSc in Teleinformatics Engineering from the Federal University of Ceará (UFC, 2007), and he graduated in Mechatronics Engineering at the Federal Center of Technological Education of Ceará (CEFETCE, 2006). He is a specialist, mainly, in Image Data Science, IoT, Machine/Deep Learning, Pattern Recognition, Automation and Control, and Robotic.