https://doi.org/10.2352/EI.2022.34.10.IPAS-367

Deep reinforcement learning approach to predict head movement in 360° videos

Tanmay Ambadkar and Pramit Mazumdar
Department of Computer Science and Engineering, Indian Institute of Information Technology Vadodara, Gujarat, India
201851018@iiitvadodara.ac.in and pramit.mazumdar@iiitvadodara.ac.in

Abstract

The popularity of 360° videos has grown immensely in the last few years. One probable reason is the availability of low-cost devices and ease in capturing them. Additionally, users have shown interest in this particular type of media due to its inherent feature of being immersive, which is completely absent in traditional 2D videos. Nowadays such powerful 360° videos have many applications such as generating various content-specific videos (gaming, knowledge, travel, sports, educational, etc.), during surgeries by medical professionals, in autonomous vehicles, etc. A typical 360° video when seen through a Head Mounted Display (HMD) gives an immersive feeling, where the viewer perceives standing within the real environment in a virtual platform. Similar to real life, at any point in time, the viewer can view only a particular region and not the entire 360° content. Viewers adopts physical movement for exploring the total 360° content. However, due to the large volume of 360° media, it faces challenges during transmission. Adaptive compression techniques have been incorporated in this regard, which is in accordance with the viewing behaviour of a viewer. Therefore, with the growing popularity and usage of 360° media, the adaptive compression methodologies are in development. One important factor in adaptive compression is the estimation of the natural field-of-view (FOV) of a viewer watching 360° content using a HMD. The FOV estimation task becomes more challenging due to the spatial displacement of the viewer with respect to the dynamically changing video content. In this work, we propose a model to estimate the FOV of a user viewing a 360° video using an HMD. This task is popularly known as the Virtual Cinematography. The proposed FOVSelectionNet is primarily based on a reinforcement learning framework. In addition to this, saliency estimation is proved to be a very powerful indicator for attention modelling. Therefore, in this proposed network we utilise a saliency indicator for driving the reward function of the reinforcement learning framework. Experiments are performed on the benchmark Pano2Vid 360° dataset, and the results are observed to be similar to human exploration.

Introduction

Omnidirectional videos or 360° videos are gaining traction recently. More due to the availability of low cost devices for easy capturing and rendering of 360° videos. Popular social networks such as Facebook and YouTube have also adopted this and are now allowing people to upload and view 360° videos (known as the user generated contents). Additionally, the traditional usage of 360° media in virtual reality (VR) applications for Head Mounted Displays (HMDs) are also growing with more sophisticated user engagement techniques related to the gaming industry. Recent HMDs provide 6 Degrees of Freedom (DOFs), which is intended for a free navigation while viewing the content. This allows a user to even walk around objects and also to look at them from all possible angles. From a gaming perspective it also allows the Gamer to dodge incoming projectiles. However, it can be noted here that even though a viewer with HMD have a free navigation with a high DOF, still one can only view a portion of the entire content at a particular timestamp. This intuitively refers to the fact that at any given timestamp, only a particular portion of the 360° video is visible.

Sophisticated cameras as used to capture 360° videos with a dimension of 360° × 180° for each frame. A viewer can freely navigate through the video either using a Virtual Reality (VR) headset or pan through using a mouse and a keyboard (2D screen based exploration). Such exploration procedure justifies the fact that only a small part of the total video is visible to the user at a given time. The portion of the total 360° scene that is visible to the user or viewer is mentioned as the natural Field-of-View (FOV) for the rest of this paper. The notion of FOV of a 360° video in a HMD can be considered similar to a real-life scenario. This is due to the fact that in real-life scenario also we are only exposed to the regions visible though the Fovea, Parafovea, Perifovea, etc. [16]. A physical movement or displacement initiates the exploration of the total scene around us in real-life scenario. Similar to this, due to the possibility of 6DOFs in modern HMD devices, subjects are free to explore the content though movement of eyes, head, trunk, and also physical movement. Estimating FOV in a 360° setting is interesting and also a challenge due to the fact that the automated model needs to correlate physical movement of subjects with the dynamic change in the video content. Intuitively attention modelling can be considered as an important indicator for estimating which FOVs a subject would look at. Attention modelling has also been successfully utilised in parallel domain of research such as medical image analysis [2, 5], autism spectrum disorder classification [1, 12], etc.

In this direction, there can be a number of driving factors which may attract viewer's attention, such as presence of human faces, a human performing a task, a fountain, etc. Therefore, presence of objects in a scene and their relation with the contextual information presented by the scene are of utmost importance to drive human attention and thus change of FOVs. Saliency estimation is proved to be a powerful approach for attention modelling. Starting from flat 2D images/ Videos, nowadays there are saliency estimation approaches that are specific to 360° content as well. In addition to this, the omnidirectional videos even with a huge popularity, faces a challenge during transmission, due to its large volume. One probable approach could be to keep the current FOV

with highest resolution and significantly compress the others, that is opting for an adaptive compression with respect to the viewing behaviour. The quality of experience models for video communication also depends on how compression is being performed [4]. Such approach could provide a seamless unbufferred experience. Popular media platforms such as Facebook[1] and Netflix[2] is working on their own visual quality metrics to estimate effect of such compression approaches with respect to visual experience of the target users viewing the content.

This work is motivated from such developments. Here we propose an estimation model that can effectively estimate the natural field-of-view of the viewers within a HMD setting. The proposed FOVSelectionNet framework effectively utilises reinforcement learning framework which also includes saliency estimation as the feature extractor. The proposed model could effectively be used to estimate the natural field-of-view of subjects in a HMD setting. Additionally, the proposed work can also help in adaptive compression of the large 360° videos. The rest of the paper is organized as follows; Section provide details of a few research works related to estimating FOV of 360° videos, Section depicts the proposed methodology, Section presents the experimental results, and finally we conclude in Section .

Related Works

Estimating the natural field-of-view for a 360° media has been extensively studied in literature [9, 22, 23]. Jabar et al. in [9] analysed the importance of scene content for viewport rendering in 360° images. Zhang et al. [23] performed head movement prediction as a sparse directed graph learning problem. The approach considers the viewer's head movement traces, saliency map for attention modelling, and the biological human head model. The Autocam model [17] uses dynamic programming to simulate human like trajectories. It captures a path over spatio-temporal glimpses to maximise capture-worthiness score and converts it to a shortest path problem. Very recently, Wang et al.[20] utilises a reinforcement learning framework to determine the FOV. It models a deep reinforcement learning agent to simulate human like behaviour with limited actions viewing an omnidirectional video. Saliency estimation is a significant indicator for attention modelling and hence can be effectively used for estimating FOV. It helps predict the human attention for a given frame. Traditional top-down and bottom-up approaches for saliency estimation utilises inherent features extracted from images [18].

Saliency was earlier estimated using traditional methods like measuring color, intensity and motion[7]. Depth field estimation performed in [3, 15] could also be used as an indicator for saliency estimation. However, saliency mostly depends upon content of a scene, and in this regard recent approaches exploits content information for estimating saliency in 360° images [13]. Newer methods use deep learning and deep neural networks to detect saliency from videos. CNNs[10] and ConvLSTMs[21] are now utilised to estimate saliency in 2D videos. Spherical CNNs [24] which can be used for omnidirectional or 360° content is also available for estimating saliency. Neural networks based models have also been developed to estimate saliency in 2D and very recently for typical 360° images [14]. SalNet360 [14] is the state-of-the-art saliency estimation framework that uses an end-to-end CNN architecture. The 360° images are subjected to various kinds of distortions. SalNet360 framework divides the omnidirectional images into undistorted patches so that the estimated saliency estimation model is not affected due to the distortion [11]. Other saliency estimation approaches using CNN includes [6, 19, 25]. In this work, we adopt the SalGan360 [6] framework as the saliency estimator.

In this work, we propose a similar framework as [20]. However, different from [20] we attempt to use an Deep Q-learning approach [8] and a different reward function. In general FOV corrections are not done for every frame, instead a series of frames are observed to finalize a FOV. We employed this strategy due to the fact that a human will not make FOV corrections for every frame. Incidentally, the viewer will observe a lot of frames and then move to another FOV. In this regard, We also experimented with different reward functions to analyse how the convergence of the agent is affected. The following section describes the methodology of the proposed FOVSelectionNet model.

Proposed Methodology

In the proposed model, we provide multi-frame inputs to the framework for predicting the next FOV. The saliency maps used in our framework are generated by adopting the saliency estimator in SalGan360 [6]. The human viewing behaviour or exploring approach can be considered as performed in 8 different directions as mentioned in [20]. The directions are as follows; North, North-East, East, South-East, South, South-West, West and North-West. In our approach, we have added Center as an additional action, thus we have 9 actions for our proposed model. Accordingly the model predicted the 9 values which are considered as the Q-values for each of the 9 actions for human exploration. The Q-values determine which action is chosen by the agent. Once an action is performed, the model receives a reward which is used to update the proposed network using back-propagation.

We select a value f which is the number of input frames for the agent. For a video with N frames, we divide it into f equal frames. Let f_t be the current frames the agent is viewing with center C_t, such that the FOV is F_t. Let S_t be the saliency maps for these frames. The input to the model will be the saliency maps S_{t+1} for frames f_{t+1}. Once an action is predicted, the model moves to that direction from the center C_t to a new center C_{t+1} with FOV F_{t+1}.

The reward function uses the FOVs F_t and F_{t+1}. It calculates the sum of the saliency values in these FOVs, denoted by SS_t and SS_{t+1} and then calculates the ratio between SS_{t+1} and SS_t.

$$Reward(R_t) = \frac{SS_{t+1}}{SS_t + 0.001} \tag{1}$$

The reward function is such that it penalises newer FOVs having a lower total saliency and encourages the model to move to regions having higher saliency. The reward is used to update the Q-values using the following equation,

$$Q(f_{t+1}, a) = R_t + \gamma max(Q(f_{t+2}, a)) \tag{2}$$

Here, we update the Q-values for state f_{t+1} using the reward R_t. This is the general update equation used in Q-learning. In

[1] $https://engineering.fb.com/2018/03/09/video-engineering/quality-assessment-of-360-video-view-sessions/$

[2] $https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12$

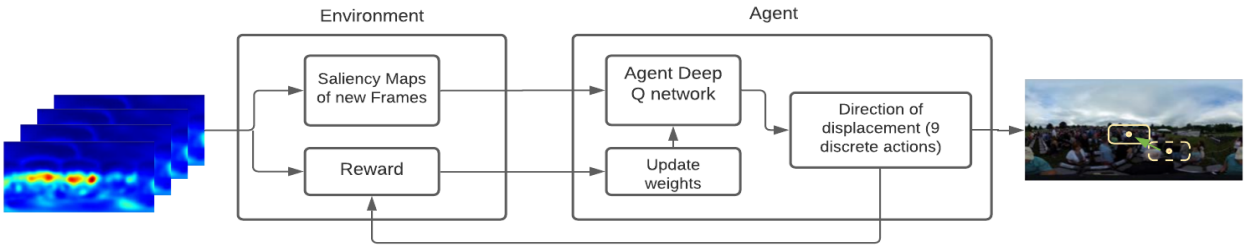

Figure 1. *Block diagram of the proposed approach.*

deep Q-learning, there are no Q tables and we use the smooth L1 loss as our criterion and the Adam optimizer for updating the weights of the neural network. A learning rate of 0.0001 is chosen for the model training.

We follow an exploration-exploitation strategy where the agent initially performs random actions to explore various strategies. With each update step, the agent reduces the exploration parameter to use the model to predict actions and get a better reward. The exploration parameter starts with a value of 1.0 and is reduced by 99.5% after every update step. The final epsilon value after which it is not reduced is 0.01. The block diagram for the proposed methodology is shown in Figure 1. Next we focus on the reward function used in our approach.

Reward function

The choice of the reward function is very important in a reinforcement learning problem. A high-valued reward will ensure that the same action is chosen again if a same/similar state is encountered. A low value reward ensures that the same action is not chosen again. We experimented with two reward functions. The first reward function is as follows,

$$Reward(R_t) = SS_{t+1} - SS_t \tag{3}$$

This reward function yields both positive and negative values. Negative values signify that the saliency of the new FOV is lower than the last FOV, meaning the model moved to a region of lower human attention. There are two possible interpretations of a negative reward. First, the agent made an incorrect prediction and moved to a region of lower saliency. Second, there is no region with higher saliency value and the agent was forced to move to a region where human attention was the highest but lower than the last FOV. An example of the reward function graph is shown in Figure 2.

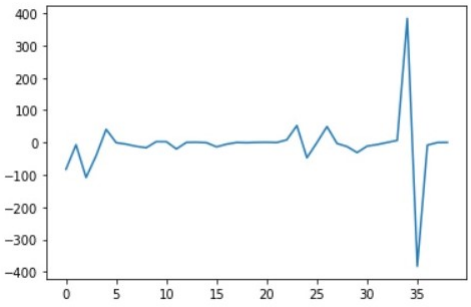

Figure 2. *Graph of the reward function 1.*

The second reward function has been illustrated in Equation 1. This reward function has a range from [0,∞). The reward is 0 when the new FOV has 0 saliency and the reward goes to infinity when the saliency in the current FOV is 0. The same interpretations can be made for the reward function as well. A reward less than 1 mean that the agent has chosen to move to regions of lower human attention. A reward greater than 1 depicts that the newer FOV has a higher saliency. Both rewards can be derived from the other. If we take the logarithm of the reward function 2, we get the first reward and if we take the exponent of reward 1, consequently, we obtain the reward function 2. In our experimental setup, the reward 2 yielded much better results than reward 1. When training the agent with reward 1, the variation in the values caused inconsistencies during the training process. Reward 2 yielded faster results, with the model choosing higher saliency valued regions more consistently. An example of reward function 2 is shown in Figure 3.

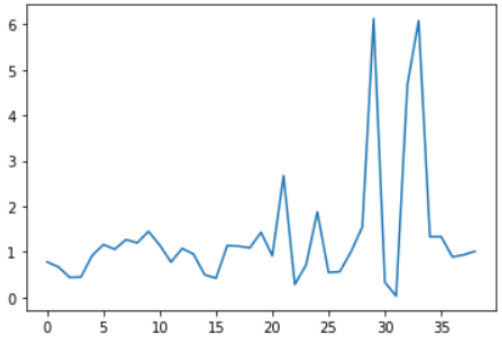

Figure 3. *Graph of reward function 2*

Next we present the FOVSelectionNet architecture along with description of each individual blocks.

Network Architecture

As mentioned earlier our architecture is provided with multi-frame inputs. Therefore, we assigned individual processing blocks (hereby named as the ConvBlocks) whose output can be merged into a single dense layer for predicting the Q-values for each action. The block diagram for processing each input saliency frame is shown in Figure 4.

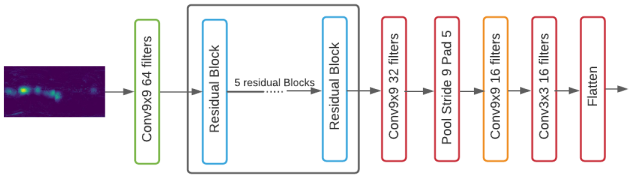

Figure 4. *A single ConvBlock*

The residual layers are used to ensure that that there is no gradient vanishing or explosion problem. It also facilitates building deeper networks. Each residual block has the structure shown in Figure 5.

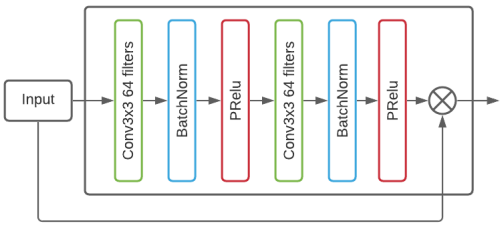

Figure 5. *A single residual block*

The proposed FOVSelectionNet is a combination of multiple ConvBlocks, as illustrated in Figure 6.

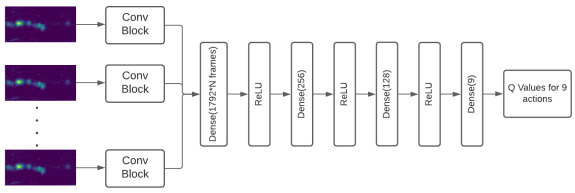

Figure 6. *Block diagram of proposed FOVSelectionNet*

Each ConvBlock output is combined into a dense layer having $1792 \times N$ dimensions, where N is the number of frames. This is used to predict the Q values for the 9 actions. The following section describes the experimental setup and obtained results.

Experimental Results

Experiments are performed on a benchmark 360° video dataset [19]. The dataset consists of multiple videos of various activities such as BMX, Skateboarding, Skating, Skiing. We experimented with different number of input frames (N) for the proposed FOVSelectionNet, where N = 1,2,4,5. It can be noted here that the presented model does not use ground truth to train the agent. The intuition behind such a choice is that exploration behaviour of an individual varies based on various factors. Thus, training on a single ground truth will limit the exploration of the entire video. The values at N=1 and 2 did not have smooth transitions. The best results for our model were recorded at N=5. At N=5, the model provided enough frames to determine the maximum human attention. The reward for the N=5 agent is shown in figure 5, where it can be observed that the model is able to find regions of higher human attention.

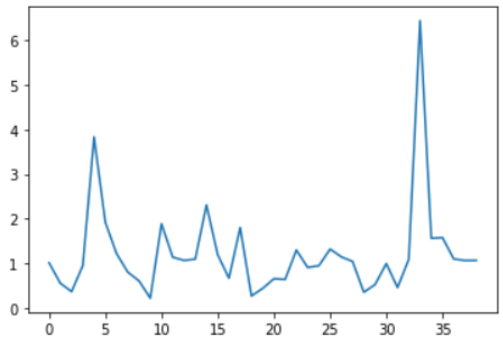

Figure 7. *Last epoch for 5 frame input.*

A few FOVs predicted by our model on various 360° videos are presented in Figure 8. The selected FOVs clearly depicts that the identified objects have the highest human attention.

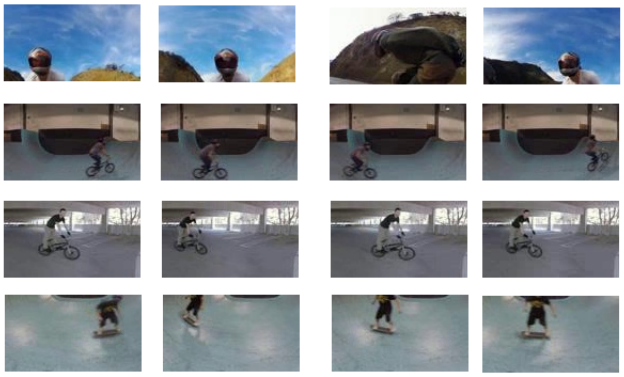

Figure 8. *FOVs selected by our proposed model.*

Conclusion

In this work, a reinforcement learning based natural field-of-view estimation model is propsed for 360° video when watched using a HMD. The model meticulously uses the saliency maps of the 360° video frames as the feature extractors. The reward function uses the saliency values and thus directs the agent towards how a human would explore the 360° content. Using multiple frames as input leads to smoother transitions from one FOV to another. Experiments are performed on benchmark dataset and the observed results justifies using the reinforcement learning learning model.

Acknowledgments

We would like to thank IIIT Vadodara for providing access to the Param Shavak server, which is a facility provided by the Gujarat Council on Science and Technology, India.

References

[1] Giuliano Arru, Pramit Mazumdar, and Federica Battisti. "Exploiting visual behaviour for autism spectrum disorder identification". In: *International Conference on Multimedia & Expo Workshops*. IEEE. 2019, pp. 637–640.

[2] Nishanth Arun et al. "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging". In: *Radiology: Artificial Intelligence* 3.6 (2021), e200267.

[3] Lucio Azzari, Federica Battisti, and Atanas Gotchev. "Comparative analysis of occlusion-filling techniques in depth image-based rendering for 3D videos". In: *Workshop on Mobile video delivery*. 2010, pp. 57–62.

[4] Federica Battisti, Marco Carli, and Pradip Paudyal. "QoS to QoE mapping model for wired/wireless video communication". In: *Euro Med Telco Conference*. IEEE. 2014, pp. 1–6.

[5] Aidan Boyd, Kevin W Bowyer, and Adam Czajka. "Human-Aided Saliency Maps Improve Generalization of Deep Learning". In: *IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 2735–2744.

[6] F. Chao et al. "Salgan360: Visual Saliency Prediction On 360 Degree Images With Generative Adversarial Networks". In: *International Conference on Multimedia Expo Workshops*. July 2018.

[7] Wen-Huang Cheng et al. "Automatic video region-of-interest determination based on user attention model". In: *International Symposium on Circuits and Systems*. IEEE. 2005, pp. 3219–3222.

[8] Matteo Gadaleta et al. "D-DASH: A deep Q-learning framework for DASH video streaming". In: *IEEE Transactions on Cognitive Communications and Networking* 3.4 (2017), pp. 703–718.

[9] Falah Jabar, João Ascenso, and Maria Paula Queluz. "Content-aware perspective projection optimization for viewport rendering of 360 images". In: *International Conference on Multimedia and Expo*. IEEE. 2019, pp. 296–301.

[10] Lai Jiang et al. "Deepvs: A deep learning based video saliency prediction approach". In: *European Conference on Computer vision*. 2018, pp. 602–617.

[11] Kamal Lamicchane, Pramit Mazumdar, and Marco Carli. "Geometric feature based approach for 360° image saliency estimation". In: *International Symposium on Image and Signal Processing and Analysis*. IEEE. 2019, pp. 228–233.

[12] Pramit Mazumdar, Giuliano Arru, and Federica Battisti. "Early detection of children with autism spectrum disorder based on visual exploration of images". In: *Signal Processing: Image Communication* 94 (2021), p. 116184.

[13] Pramit Mazumdar and Federica Battisti. "A content-based approach for saliency estimation in 360 images". In: *International Conference on Image Processing*. IEEE. 2019, pp. 3197–3201.

[14] Rafael Monroy et al. "Salnet360: Saliency maps for omni-directional images with cnn". In: *Signal Processing: Image Communication* 69 (2018), pp. 26–34.

[15] Alessandro Neri, Marco Carli, and Federica Battisti. "A multi-resolution approach to depth field estimation in dense image arrays". In: *International Conference on Image Processing*. IEEE. 2015, pp. 3358–3362.

[16] Hans Strasburger, Ingo Rentschler, and Martin Jüttner. "Peripheral vision and pattern recognition: A review". In: *Journal of vision* 11.5 (2011), pp. 13–13.

[17] Yu-Chuan Su, Dinesh Jayaraman, and Kristen Grauman. "Pano2Vid: Automatic Cinematography for Watching 360° Videos". In: *Asian Conference on Computer Vision*. Springer. 2016, pp. 154–171.

[18] Inam Ullah et al. "A brief survey of visual saliency detection". In: *Multimedia Tools and Applications* 79.45 (2020), pp. 34605–34645.

[19] C. H. Vo et al. "Saliency Prediction for 360-degree Video". In: *International Conference on Green Technology and Sustainable Development*. 2020, pp. 442–448.

[20] Jianyi Wang et al. "Attention-based Deep Reinforcement Learning for Virtual Cinematography of 360° Videos". In: *IEEE Transactions on Multimedia* (2020).

[21] SHI Xingjian et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting". In: *Advances in Neural Information Processing Systems*. 2015, pp. 802–810.

[22] Tan Xu, Bo Han, and Feng Qian. "Analyzing viewport prediction under different VR interactions". In: *International Conference on Emerging Networking Experiments And Technologies*. 2019, pp. 165–171.

[23] Xue Zhang et al. "Graph Learning Based Head Movement Prediction for Interactive 360 Video Streaming". In: *IEEE Transactions on Image Processing* 30 (2021), pp. 4622–4636.

[24] Ziheng Zhang et al. "Saliency detection in 360 videos". In: *European conference on computer vision*. 2018, pp. 488–503.

[25] Dandan Zhu et al. "Saliency prediction on omnidirectional images with attention-aware feature fusion network". In: *Applied Intelligence* (2021), pp. 1–14.

Author Biography

Tanmay Ambadkar *is pursuing his B.Tech in Computer Science from Indian Institute of Information Technology Vadodara, India and is working as an RD Intern at Siemens Technology and Services Pvt. Ltd. He specializes in Deep Learning in computer vision problems and reinforcement learning.*

Pramit Mazumdar *received Ph.D. degree from the National Institute of Technology Rourkela, India. He is an Assistant Professor at the Department of Computer Science and Engineering, Indian Institute of Information Technology Vadodara, India. He was a Visiting Researcher with the University of Roma Tre, Rome, Italy, from 2018 to 2021. His research interests are in the area of digital signal and image processing with applications to multimedia quality evaluation and communications.*