IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 August 2022, accepted 6 September 2022, date of publication 12 September 2022,
date of current version 22 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3206045

RESEARCH ARTICLE

Eye Tracking, Saliency Modeling and Human Feedback Descriptor Driven Robust Region-of-Interest Determination Technique

MANORANJAN PAUL[1], (Senior Member, IEEE), PALLAB KANTI PODDER[1], (Member, IEEE), AND MD. RIAD HASSAN[2], (Student Member, IEEE)

[1]School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia
[2]Institute of Information and Communication Technology, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh

Corresponding author: Manoranjan Paul (mpaul@csu.edu.au)

This work was supported in part by the Australian Research Council through the Discovery Projects under Grant DP190102574.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Human Research Ethics Committee of Charles Sturt university under Approval No. 2015/124.

ABSTRACT The *Region of interest* (ROI) analysis is widely used in image analytics, video coding, computer graphics, computer vision, medical imaging, nuclear medicine, computer tomography and many other areas in medical applications. This ROI determination process using subjective method (e.g. using human vision) often differ from the objective ones (e.g. using mathematical modelling). However, there is no existing method in the literature that could provide a single decision when both methods' ROI data is available. To address this limitation, a robust algorithm is developed by combining the human eye tracking (subjective) and the graph-based visual saliency modelling (objective) information to determine a more realistic ROI for a scene. To carry out this process, in one hand, several different independent human visual saliency factors such as pupil size, pupil dilation, central tendency, fixation pattern, and gaze plot for a group of twenty-two participants are collected by applying on a set of publicly available eighteen video sequences. On the other hand, the features of *Graph based visual saliency* (GBVS) highlights conspicuity in the scene. Gleaned from these two pieces of information, the proposed algorithm determines the final ROI based on some heuristics. Experimental results show that for a wide range of video sequences and compared to the existing deep learning based (MxSalNet) and depth pixel (DP) based ROI, the proposed ROI is more consistent to the benchmark ROI, which was previously decided by a group of video coding experts. As the subjective and objective options frequently create an ambiguity to reach a single decision on ROI, the proposed algorithm could determine an ultimate decision, which is eventually validated by experts' opinion.

INDEX TERMS Eye tracking, expert opinion, GBVS, region-of-interest, visual saliency.

I. INTRODUCTION

Measurement of eye movements has been extensively employed in visual attention, *region of interest* (ROI) determination and perception modelling including image and video analytics [1], mammography [2], classroom education [3] and many more [4]. The ROI analysis is seemingly used in image/video analytics, computer graphics, computer vision, medical imaging, nuclear medicine, computer tomography and many more areas in medical applications [5], [6], [7], [8]. This ROI determination process could be carried out by subjective and objective methods. The subjective estimation involves human in the process, such as mean opinion score (MOS) and eye maneuver (EMAN) for quality detection, which is very similar to the use of eye-tracking in saliency/ROI determination. The Objectives methods rely more on mathematical and statistical modelling such as the Statistical similarity (SSIM) for quality and graph based

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

visual saliency (GBVS) for ROI detection. The subjective studies could yield valuable data to evaluate the performance of objective methods towards aiming the goal of matching human perception. The subjective estimation employs the utilization of human visual attention and its parameters such as pupil size, pupil dilation, timestamp data, central tendency, fixation pattern, and gaze plot [9], [10], [11], [12]. The objective estimation, for example, Graph based visual saliency (GBVS), on the other hand, focuses on colour contrast, brightness and motion on spatiotemporal domain [13]. The objectively determined ROI is widely used for its simplicity of use, however, the human visual system is the ultimate assessor of determining the insights of a video and there is a growing demand of ROI determination using subjective method.

Human visual attention regions can be recorded using remote screen or head-mounted eye tracking system while watching a given video clip. Moreover, the visual perception these days can also be captured and estimated by employing the software-based gaze estimation simulator where the device itself is no longer needed [14]. In the literature, a number of research works have been proposed based on visual data analysis to predict gaze region in image and video [15], [16], [17], [18], [19]. Most of the contributions presented here use some statistical correlation to determine fixation mapping, saliency-based visual prediction, object tracking and human attention in a scene. However, literature shows that more accurate approach to determine the actual gaze locations is to use a gaze-tracking device (e.g. eye tracker) [20].

Eye tracker recorded video data has been exploited by many researchers to serve several real-life purposes such as video summarization [21], [22], [23], cognitive model generation and visual model fixation [24], [25], [26], [27], [28], [29], [30]. In the work of Zheng *et al.* [31], a fusion model was presented to improve the performance of emotion recognition by combining human pupillary responses (by using pupil diameter only) collected from eye tracker and electroencephalograph (EEG) signals. Hadizadeh *et al.* [32] provided an eye tracking database based on first and second viewings of fifteen individuals for twelve uncompressed standard video sequences. They compare average distance between the first and second viewings in pixels as well as percentage of frame diagonals and analyse the fluctuations of two viewings especially for Foreman standard video sequence. In addition, they compared Itti-Koch-Niebur (IKN) [19] and Itti-Baldi (IB) [18] visual attention models and while observing that IKN showed better accuracy than IB based on the eye tracker data as the benchmark ground truth. Jia *et al.* [33] proposed a no-reference video quality assessment algorithm based on eye tracking data for four different videos.

Dodge *et al.* [34] proposed a Visual Saliency Prediction Using a Mixture of Deep Neural Networks (*MxSalNet*) where the final saliency map is computed as a weighted mixture of networks. Zhou *et al.* [35] proposed depth pixel (DP) based ROI by characterizing the depth image fusion. Zhang *et al.* [36] proposed a co-saliency based ROI detection system by using a manifold technique where image correlation, energy function and fusion weights are considered. This is a mathematical model-based method to represent ROI, which lacks the limitation of involving human opinion to determine or verify final ROI. Ma *et al.* [37] proposed a ROI extraction model based on unsupervised cross-domain adaptation. This process incurs high computation complexity as it goes through some prior learning process and still focuses on mathematical modelling and lacks visual perception.

Sun *et al.* [38] proposed a system that determines ROI from eye tracker data using a monocular camera. To identify, ROI, He *et al.* [39] proposed Fourier Transform based graph signal processing and clustering system to classify data samples from noisy eye tracker data. Their tested results claim competitive clustering accuracy of ROI; however, this process suffers from determining a single ROI where multiple number of ROIs exist in a complex scene. Others clustering approaches have been used to determine ROI, such as Density-based spatial clustering of applications with noise (DB-SCAN) [40], k-means and distance threshold [41], [42], Distance-Threshold Identification (IDT) [43] and Mean-shift [44]. They try to reduce noise and determine the ROI from eye tracker data.

Therefore, our motivation is to draw a close comparison between eye tracking and GBVS generated salient point, acquire knowledge of their similarity-divergence relationship, employ a number of visual sensitive features to highlight conspicuous region in the scene, apply some heuristics to determine the final ROI gleaned from subjective and objective information, and finally compare it to the benchmark ROI determined by experts' opinion to revive a more realistic ROI of a scene. Beside this, to demonstrate consistency, the proposed method is compared with the recent deep learning-based ROI and depth pixel-based ROI estimation methods. This work can be applied in the areas of video compression, medical image analysis, image segmentation and many more.

The main contribution of this paper can be summarized as follows:

(i) Carry out a comprehensive analysis on eye-tracking data and associated parameters for visual saliency modelling.

(ii) Investigate similarity-divergence relationship by making a close comparison between eye-tracking (i.e. subjective) and graph-based visual saliency (i.e. objective) modelling information to develop a robust ROI.

(iii) Mathematically analyze the parameters to fix in-focus region by analyzing eighteen videos seen by twenty-two people.

(iv) Develop an algorithm to determine a more realistic ROI of a scene when both subjective and objective information are available.

(v) Compare the proposed algorithm with the recent deep neural network and depth pixel based ROI approaches.

(vi) Incorporate a group of video coding experts' opinion to justify and validate finally decided ROI.

The remainder of the paper is organized as follows: Section-II focuses on proposed experimental set-up; Section-III presents the experimental detail; the experimental results are broadly discussed in Section-IV, while Section-V concludes the paper.

II. PROPOSED EXPERIMENTAL SET-UP

Voluntary participants (twenty male and two female) were recruited in the university through an open invitation disseminated through emails and notice board posters which included a detailed 'Participant Information Sheet' about the project. They had normal or corrected-to-normal vision and did not suffer from any medical condition that might adversely influence our project. The participants remain anonymous, i.e. they were known as person 1, person 2 and so on. Only their age and gender were recorded in order to identify any possible pattern emergence. They fall within the 20-45 age band who are undergraduate/postgraduate students, PhD students, and lecturers of the university. Technical details of Tobii eye tracker working, and safety standard were also conveyed. Publicly available and widely used eighteen video clips were sequentially shown to the participants at twice their normal size so that they could cover over 80% of the screen. The videos were demonstrated at a fixed order with a 2-s pause in between.

The Tobii eye tracker (ET) uses a computer software known as Tobii Studio 2.0 to record information associated with eye gaze data i.e. the particular points where the users' focused on, their pupil sizes, eye blinking pattern during the experiments. There is no physical contact between the participants and the device. For displaying video 24-Inch full HD (1920 x 1080 pixel) monitor is used. Before starting experiment, eye tracker needed calibration with participant eye and the video display monitor. This calibration goes through a proper mapping process which eventually negate other factors, such as participant's sitting position, distance from the monitor, and monitor size. As, the device collects data every 16.5 milliseconds on average, it can collect all significant gaze point. Thus, every chunk of second is important in data collection [45], [46]. Moreover, this system removes all scattered data by considering active vision [47]. The short movies employed are of different lengths (4 to 9 seconds), have common intermediate format (CIF- 352 × 288 resolution), 30 frames per second and in 4:2:0 YUV format and are well known to the video research community.

At a later time, a second group to people (6 participants) who are video experts and have detail knowledge about compression, transmission, and processing of multimedia watched the eighteen videos separately to opine about the most significant or attention point in each clip. The rationale behind selecting is its simplicity, having an appearance of ground truth, seemingly worthy of approval and its natural parallelism property. Different studies show that it performs better compared to other saliency models [48].

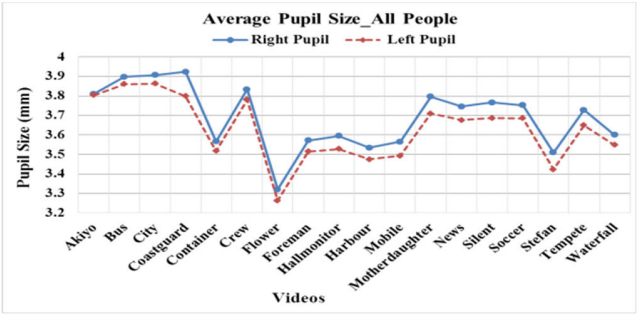


FIGURE 1. Illustration of average pupil size for all participants over eighteen videos.

III. EXPERIMENTAL DETAIL

The task of calculating the ROI is executed based on the parameters entitled (i) Average pupil size (ii) Fixation pattern (iii) Eye tracker recorded gaze point [47] (iv) Experts' opinion point and (v) GBVS salient point. We denote these parameters as PSA, FP, GPA, OPE, VSP respectively where the pupil sizes are measured in millimetre (mm), fixation patterns in percentage (%) and the distances in pixels. The eighteen sequences used in this experiment are Akiyo, Bus, City, Coastguard, Container, Crew, Flower, Foreman, Hallmonitor, Harbour, Mobile, Motherdaughter, News, Silent, Soccer, Stefan, Tempete, and Waterfall with the resolution of 352 × 288.

A. AVERAGE PUPIL SIZE

It is noticed from **FIGURE 1** that for almost all the sequences used in this experiment, regardless of considering its duration and emotional sensitivity, the right eye pupil sizes are always greater than left ones for all participants. It is also noticed that almost in all videos, the average left and right pupil size of each individual participant is equal or greater than 3.5 mm. The normal pupil size tends to range between 2.0 and 5.0 mm depending on the lighting. As the effect of lighting was not taken into account in this experiment, the recorded pupil sizes would be suitable to capture relevant information while watching videos [9].

B. FIXATION PATTERNS

Eye blinking pattern in this work has been broken down into two different phases- the fixation and the unclassified. Fixation is the period determined by the visual gaze on a single location. In contract, unclassified visual data indicate the time when participants' eye traversal is not recognized by the eye tracker due to the closure of eyes, movement of head, or scattered vision outside of the visual display region set during the set-up of the experiment. Please note that the overall calculated unclassified data was less than 3% in the entire experiment. The collected data (see **FIGURE 2**) indicate participants had an average fixation rate of 97%. [49].

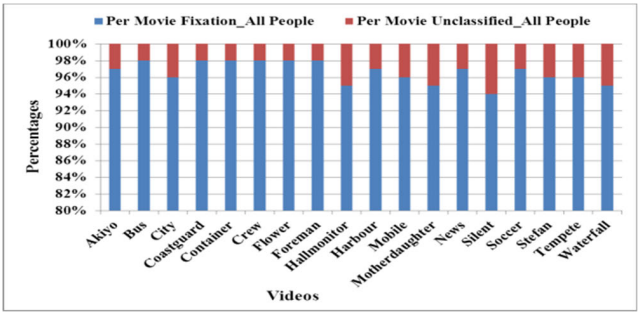


FIGURE 2. Video wise average fixation percentage for all participants.

C. EYE TRACKER RECORDED GAZE POINT

Data captured from the eye tracker can provide the individual's gaze points for all frames. At first the eye tracker gaze locations data that was collected for group 1 (12 participants) was averaged person wise and video wise separately to locate the most attentive points. An example of GPA for Akiyo video is calculated by averaging 300 fixation points. In this way, the fixation points of twelve participants over eighteen videos are calculated. Eight random videos from experimental eighteen videos are presented in **FIGURE 3** where red marked point is the gazed point of different participant.

D. EXPERTS OPINION POINT

A second group to people (six in our experiment and separate from the participants' set) who are video experts and have detail knowledge about compression, transmission, and processing of multimedia watched the eighteen videos separately to opine about the most significant or attention point in each clip. Expert opinion point is represented in **FIGURE 3** where black marked point is expert opinion point.

E. GBVS SALIENT POINT

GBVS is applied on the difference between two successive frames for the entire duration of the video to generate a number of salient points and the final salient point is calculated by averaging the 20% maximum value of the GBVS generated salient points. The GBVS saliency along with GBVS salient point of eight videos from our experiment is shown in **FIGURE 4** where GBVS salient point is marked with red rectangle which cover 20% area around of salient point.

F. MxSalNet SALIENT POINT

MxSalNet is a deep neural network based visual saliency prediction method [34]. We apply this method for predicting saliency of images and finally we determine the MxSalNet salient point by averaging first 20% of maximum values from *MxSalNet* generated saliency. The *MxSalNet* based saliency and salient point of eight different videos are depicted in **FIGURE 5**. Here, the salient point is marked with red color box which covers 20% surrounding of salient point.

FIGURE 3. Human eye tracked recorder gaze points (Red marked) and expert opinion (Black marked) in eight randomly selected videos from experimental videos.

G. DEPTH PIXEL BASED SALIENT POINT

Depth pixel information of image is used to determine the ground pane and filter out the probable non-ROI points from image. Then it finds the candidate portions for ROI of image using window sliding method. Finally among the candidate portions, it selects one ROI using candidate option filtering method [35]. We apply this method and carry out the salient point by averaging maximum 20% of depth pixel based salient points. Salient points of four different videos are illustrated in **FIGURE 6**. Here, the boxes (Red color for candidate, green color for final) of salient points are marked which cover 20% around area of salient point.

H. DETERMINATION OF ROI FROM ETRD

FIGURE 7 demonstrates the ROI determination process from the average gaze locations of twelve participants for the City video. In **FIGURE 7** (b), the second participant's average gaze location for the first half and second half is presented pictorially. The reason of selecting the second person is his highest concentration (100%) to this video. The bottom-centre and top-left-centre squares show the way of viewing for the first and second half of video duration

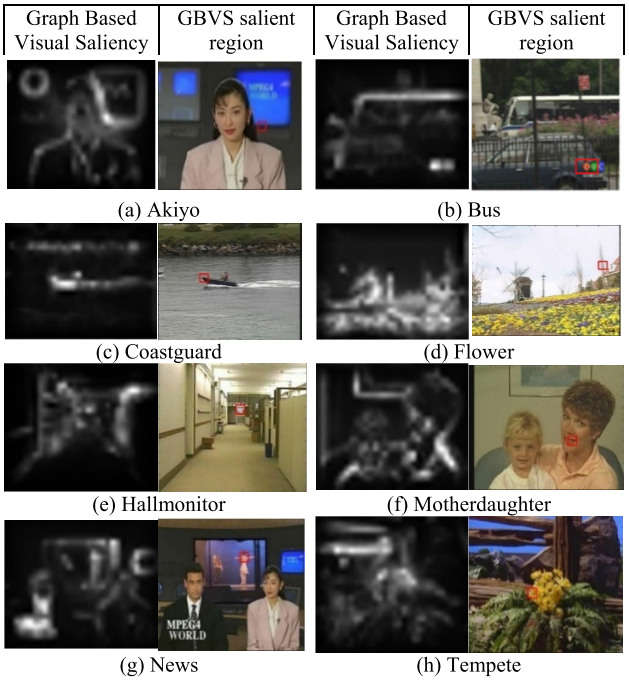


FIGURE 4. GBVS saliency and GBVS salient region.

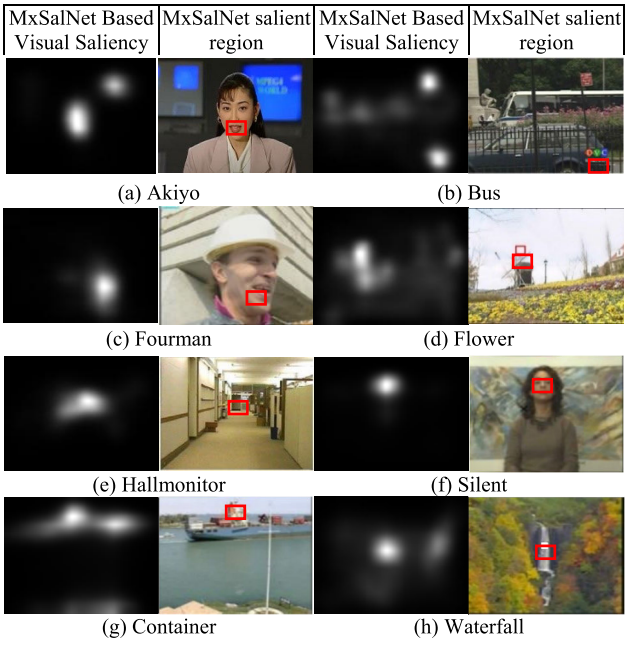


FIGURE 5. MxSalNet based saliency and MxSalNet salient point and associated region.

respectively for this participant. Then the average gaze locations for a single participant and all participants are calculated and shown in **FIGURE 7** (c)-(d). A fixation point is converted into the ROI by considering 20% surrounding pixels of that point for better visualization.

Now, we analyse every video in the context of human visual system data, which is captured using eye tracker (repre-

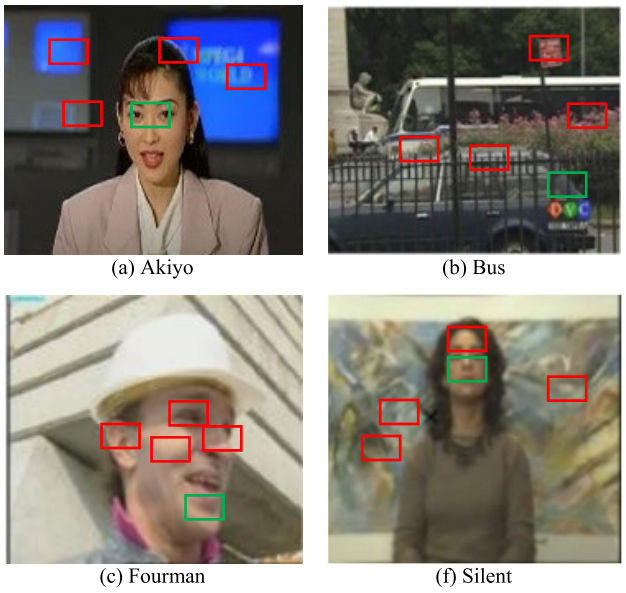


FIGURE 6. Depth pixel-based saliency point. Here red and green marked points are candidate, but green marked area is finally selected saliency region.

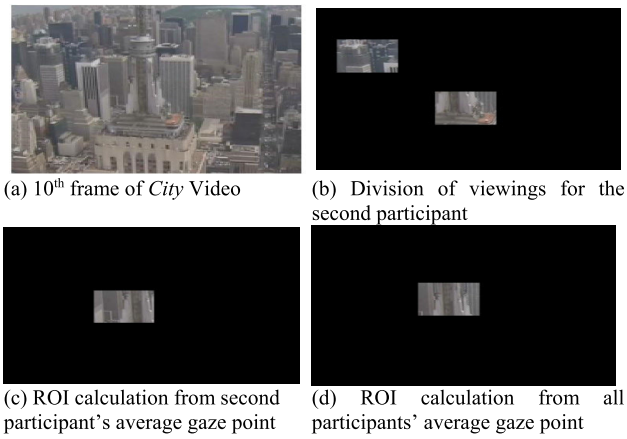
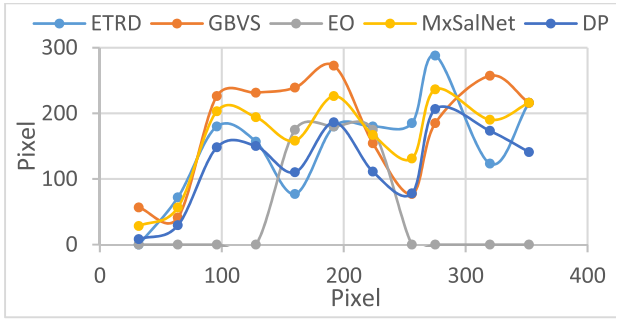


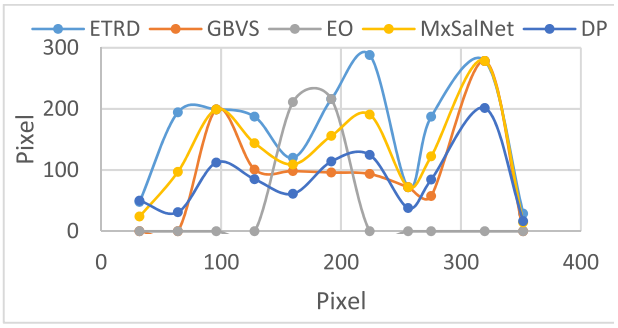
FIGURE 7. Determination of ROI from the average gaze locations.

sented as ETRD), mathematical model based GBVS, expert opinion based EO, Deep learning based MxSalNet, and the Depth pixel-based DP saliency model. From **FIGURE 8**, is observed that the human visual system (i.e. subjective estimation) for most of the videos is closely related to the expert opinion while, MxSalNet, DP, and the GBVS do not always become identical with the subjective ones because of its operational dependency on high-contrast, object motion, brightness, and resolution.

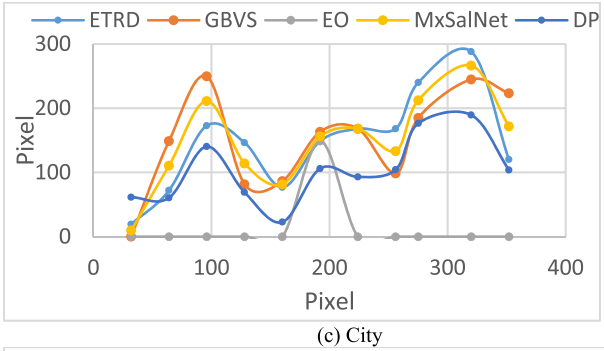
As human visual system data is closely related to expert opinion as well as this group has specialization on video analysis, coding, compression, quality, image processing, we take expert opinion as ground of our analysis. From the analysis of a particular frame of videos, we observe that there is a good co-relation between Human visual system and Expert opinion, while DP, MxSalNet, and GBVS are far from Expert



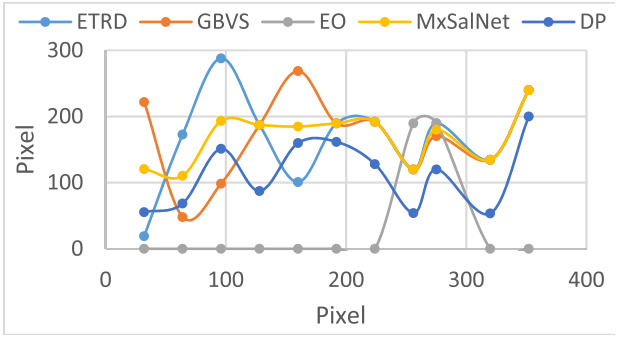
(a) Akiyo



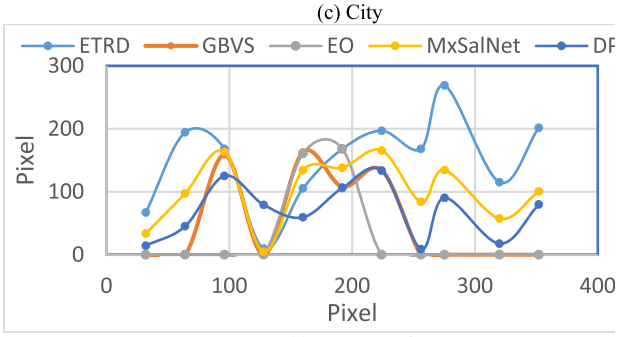
(b) Bus



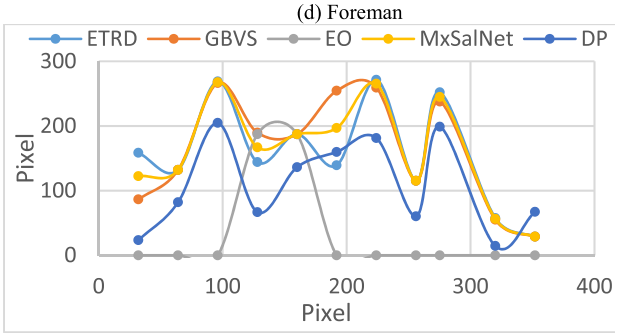
(c) City



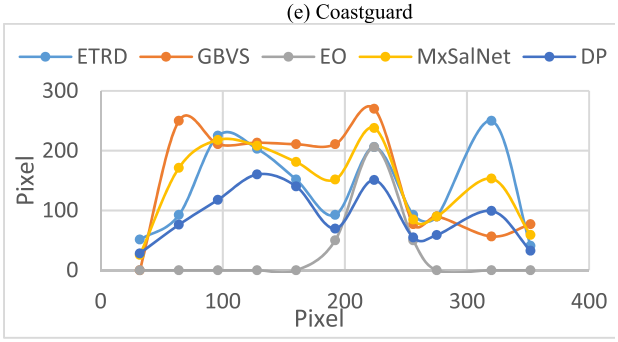
(d) Foreman



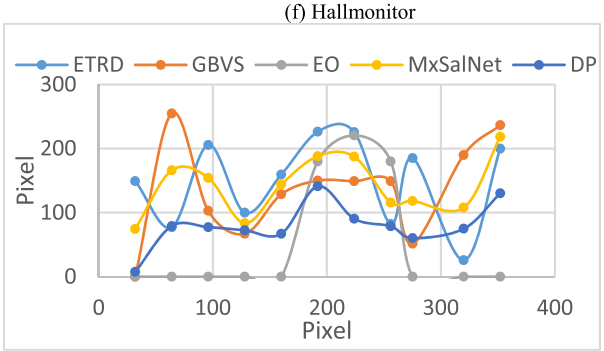
(e) Coastguard



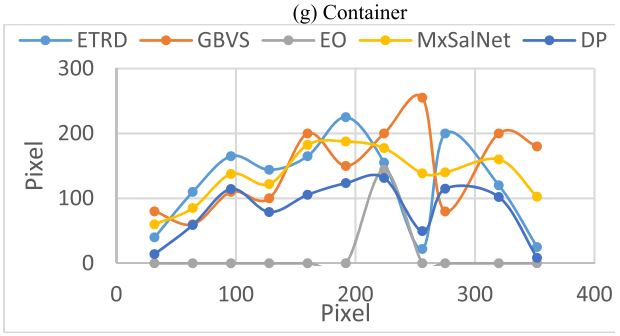
(f) Hallmonitor



(g) Container



(h) Harbour



(i) Crew

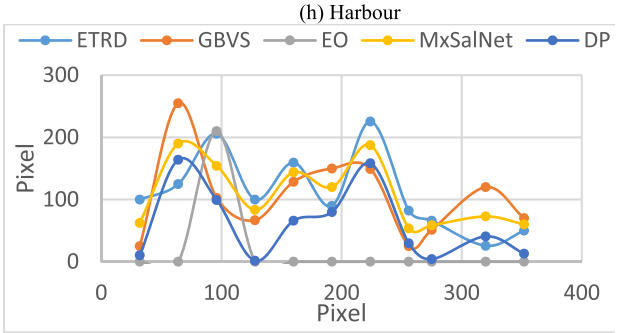
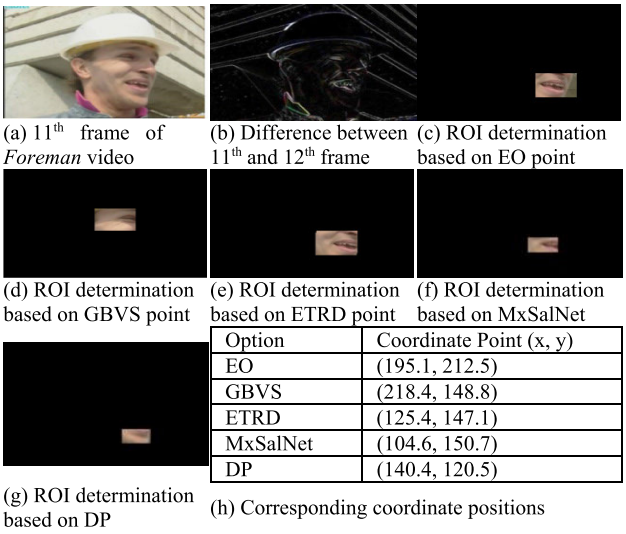


FIGURE 8. Video wise comparison among Eye Tracked Recorded Data (ETRD), Graph Based Visual Saliency (GBVS) and Expert Opinion (EO), Deep learning (MxSalNet), Depth pixel (DP) based region of interest.

FIGURE 8. *(Continued.)* **Video wise comparison among Eye Tracked Recorded Data (ETRD), Graph Based Visual Saliency (GBVS) and Expert Opinion (EO), Deep learning (MxSalNet), Depth pixel (DP) based region of interest.**

opinion. Thus, it is obvious that software-based ROI is not always steadfast to define actual ROI of human.

FIGURE 9 – FIGURE 12 present the ROI determination process from the experts' opinion (EO), GBVS eye tracker recorded data (ETRD) generated gaze points, deep learning based MxSalNet, Depth pixel base saliency for Foreman, Bus, Soccer and Stefan respectively. **FIGURE 9** (c) and **FIGURE 9** (e) show ROI pattern perceived from EO and ETRD based data, however, that differs from GBVS determined ROI (see **FIGURE 9** (d)). Here, **FIGURE 9** (f) and **FIGURE 9** (g) represent *MxSalNet* deep learning ROI and depth pixel-based ROI for Foreman video. Their corresponding ROI based coordinates are provided in **FIGURE 9** (h). Though the GBVS, Depth pixel, Deep Learning based model

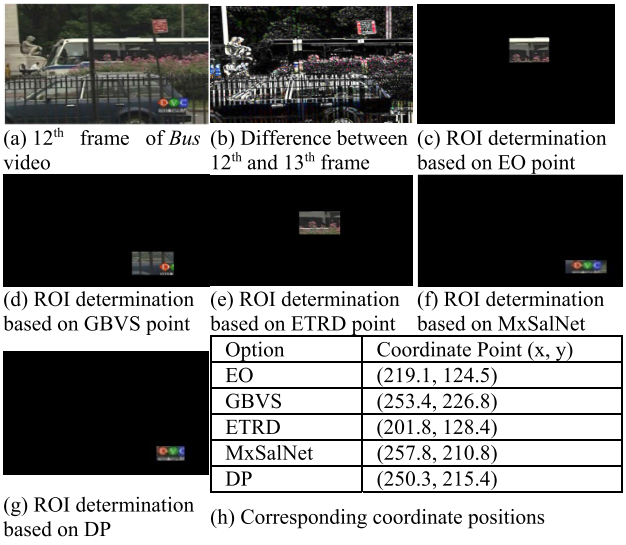


(a) 11th frame of *Foreman* video
(b) Difference between 11th and 12th frame
(c) ROI determination based on EO point
(d) ROI determination based on GBVS point
(e) ROI determination based on ETRD point
(f) ROI determination based on MxSalNet
(g) ROI determination based on DP

Option	Coordinate Point (x, y)
EO	(195.1, 212.5)
GBVS	(218.4, 148.8)
ETRD	(125.4, 147.1)
MxSalNet	(104.6, 150.7)
DP	(140.4, 120.5)

(h) Corresponding coordinate positions

FIGURE 9. Determination of ROI from the EO, GBVS, ETRD, MxSalNet and DP generated gaze points for Foreman video.

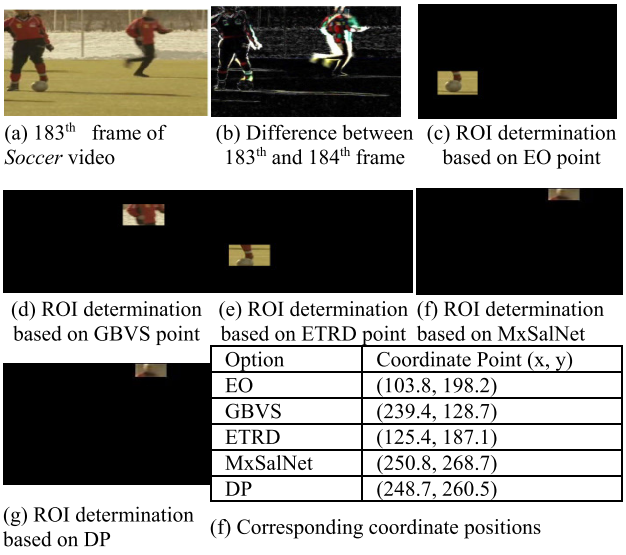


(a) 12th frame of *Bus* video
(b) Difference between 12th and 13th frame
(c) ROI determination based on EO point
(d) ROI determination based on GBVS point
(e) ROI determination based on ETRD point
(f) ROI determination based on MxSalNet
(g) ROI determination based on DP

Option	Coordinate Point (x, y)
EO	(219.1, 124.5)
GBVS	(253.4, 226.8)
ETRD	(201.8, 128.4)
MxSalNet	(257.8, 210.8)
DP	(250.3, 215.4)

(h) Corresponding coordinate positions

FIGURE 10. Determination of ROI from the EO, GBVS, ETRD, MxSalNet and DP generated gaze points for bus video.

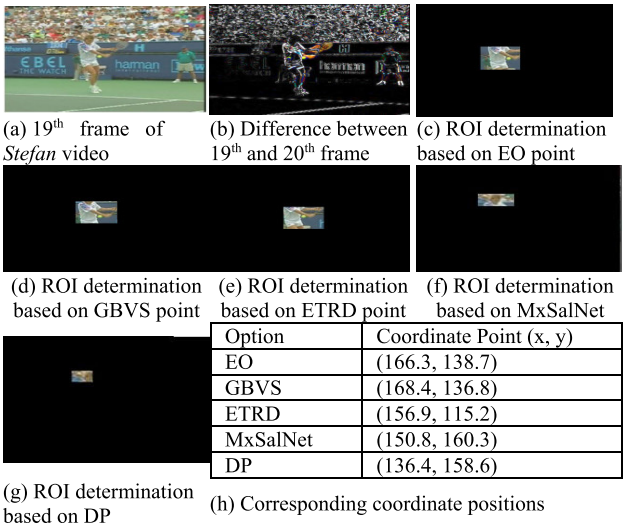


(a) 183th frame of *Soccer* video
(b) Difference between 183th and 184th frame
(c) ROI determination based on EO point
(d) ROI determination based on GBVS point
(e) ROI determination based on ETRD point
(f) ROI determination based on MxSalNet
(g) ROI determination based on DP

Option	Coordinate Point (x, y)
EO	(103.8, 198.2)
GBVS	(239.4, 128.7)
ETRD	(125.4, 187.1)
MxSalNet	(250.8, 268.7)
DP	(248.7, 260.5)

(f) Corresponding coordinate positions

FIGURE 11. Determination of ROI from the EO, GBVS, ETRD, MxSalNet and DP generated gaze points for soccer video.



(a) 19th frame of *Stefan* video
(b) Difference between 19th and 20th frame
(c) ROI determination based on EO point
(d) ROI determination based on GBVS point
(e) ROI determination based on ETRD point
(f) ROI determination based on MxSalNet
(g) ROI determination based on DP

Option	Coordinate Point (x, y)
EO	(166.3, 138.7)
GBVS	(168.4, 136.8)
ETRD	(156.9, 115.2)
MxSalNet	(150.8, 160.3)
DP	(136.4, 158.6)

(h) Corresponding coordinate positions

FIGURE 12. Determination of ROI from the EO, GBVS, ETRD, MxSalNet and DP generated gaze points for Stefan video.

claims that they can predict human attention points efficiently, our experimental data does not always effectively reflect that. The reason might be due to the way this saliency model works where salient areas in an image may be considered with high motion, resolution, colour region. In the video clip Bus, a long single-decker bus is seen moving in and around the centre of the screen, while GBVS, Deep learning based saliency, depth pixel based ROI are concentrating to three colored dots (red, green, blue) visible on the bottom right of **FIGURE 10** (d, f, g).

Another example could be provided with the Soccer video clip where several soccer players are seen practicing with a football. The players are wearing colourful jerseys but the colour of the ball was somewhat not that bright. It is noticed that the GBVS picked up the coloured regions as the most significant points shown in **FIGURE 11** (d). Deep learning based MxSalNet saliency, depth pixel-based ROI are concentrating in human face **FIGURE 11** (f, g). In contrast, the video experts and other participants' attention points were primarily focused nearby regions of the football. For the Stefan video in **FIGURE 12**, all the three estimators (GBVS, ETRD, EO) opine almost to the same points. Here, as previous deep learning based MxSalNet saliency, depth pixel-based ROI are concentrating in human face **FIGURE 12** (f, g). However, it is noticed from the exemplified videos that in most cases, GBVS predicts the salient points either concentrating to the centre or any other coloured regions. Interestingly, video content-based points obtained by EO and ETRD have an

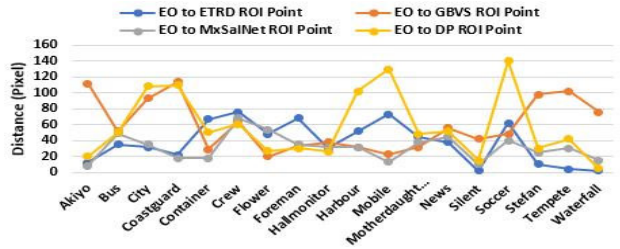


FIGURE 13. Distance from EO point to ETRD, MxSalNet, DP, and GBVS salient points calculated for eighteen videos.

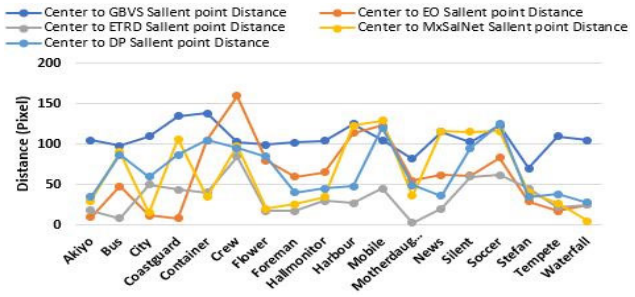


FIGURE 14. Video wise average pixel, Experts' opinion, GBVS, Deep Learning (MxSalNet) and Depth Pixel based ROI distances from the videos' centers.

identical relationship of selecting different attention points in different video clips.

It could be seen from **FIGURE 13** that the distance from video experts' opinions i.e. human feedback, of the videos to the most significant salient points obtained from ETRD and GBVS differ for most of the video clips. The figure reveals that for eleven videos, ETRD shows the minimum distance with EO points, while for seven videos GBVS shows the minimum distance with EO. We further calculate the pixel distances from the video centres to all the salient points determined by three estimators for all videos as shown in **FIGURE 14**. For most of the videos the calculated distance from the centre to GBVS point is the maximum compared to ETRD or EO points.

For further comparison, we apply three tyring approaches to the calculated distances compared to the centre as shown in **FIGURE 15**.

The tyre closest to the centre is the first tyre, then second and so on. It is noticed that both EO and ETRD mainly focus on the first and second tyres. The reason may be image capturing technique where all the sensitive and salient points in the images are captured by keeping those points at the centre of the camera screen. Being unaware of the centre sensitivity the GBVS rather considers high motion and resolution, or coloured regions in the videos.

We develop a mathematical model for EO, ETRD and GBVS focusing tyre concept from eye tracker data i.e. pupil size $\forall$, fixation ∂, distance from centre of gaze location μ and distance from salient point of gaze location φ. Focusing tyre Φ from eye tracker data will be

$$\Phi_{\text{ETRD}} = (\forall\partial)^{\frac{1}{\gamma}}\left(\frac{1}{\mu}+\frac{1}{\varphi}\right)^{\gamma} \tag{1}$$

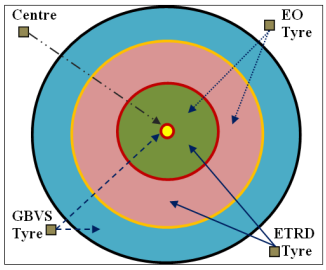


FIGURE 15. Apply tyring concept on an image to determine center sensitivity.

$$\Phi_{\text{EO}} = \frac{1}{5}\left[\forall\partial\left(\frac{1}{\mu}+\frac{1}{\varphi}\right)^{2\gamma}\right] \tag{2}$$

$$\Phi_{\text{GBVS}} = (\forall\partial)^{\frac{\gamma}{2}}\left(\frac{1}{\mu}+\frac{1}{\varphi}\right)^{\frac{\gamma}{10}} \tag{3}$$

Here, $\gamma = 0.125$. The value of Φ range is 0 to 1 where 0 consider as centre of the image and 1 is the last outer focusing tyre. When we provide the eye tracker data i.e. $\forall$, ∂, μ and φ, it provides the tyre of focusing tyre Φ of ETRD, EO and GBVS for that video.

Algorithm 1 Determination of a Realistic ROI From Graph Based Visual Saliency and Eye Tracker Recorded Information

1. Initialize,
   $\epsilon \leftarrow$ ROI based on ETRD
   $\beth \leftarrow$ ROI based on GBVS
   $f \leftarrow$ Fixation in millisecond
   (x, y) $\leftarrow$ n-th frame size.
2. Calculate,
   Centre of the frame, $\alpha \leftarrow (x \div 2, y \div 2)$
   Motion, $\exists \leftarrow$ n th frame – (n-1) th frame
   Motion pixels count, $\mho \leftarrow$ count_non_zero_pixel($\exists$)
   Motion compared to frame size $\Theta \leftarrow \mho \div (\text{x} \times \text{y})$
   Distance centre to ETRD-ROI, $\rho \leftarrow |\alpha - \epsilon|$
   Distance centre to GBVS-ROI, $= \tau \leftarrow |\alpha - \epsilon|$
3. IF $\Theta \leq 0.05$ & $f \geq 50$:
   ROI $\leftarrow \epsilon$
   ELSE IF $0.05 < \Theta \leq 0.10$:
   ROI $\leftarrow \frac{\epsilon+\alpha}{2}$
   ELSE:
   IF $\tau - \rho > 80$:
   ROI $\leftarrow \frac{\epsilon+3}{2}$
   ELSE:
   ROI $\leftarrow \frac{\epsilon+\alpha}{2}$
   END IF

END IF

Visual observant areas may not always stick to the center. For any sports video like *Soccer*, position of the ball always keeps changing while game is in continuation. Hence the focal point needs to be considered; may be the surroundings of the ball. Like GBVS, high motion and brightness dominated areas, like red, yellow and green light symbols also provide significant information which may also be considered

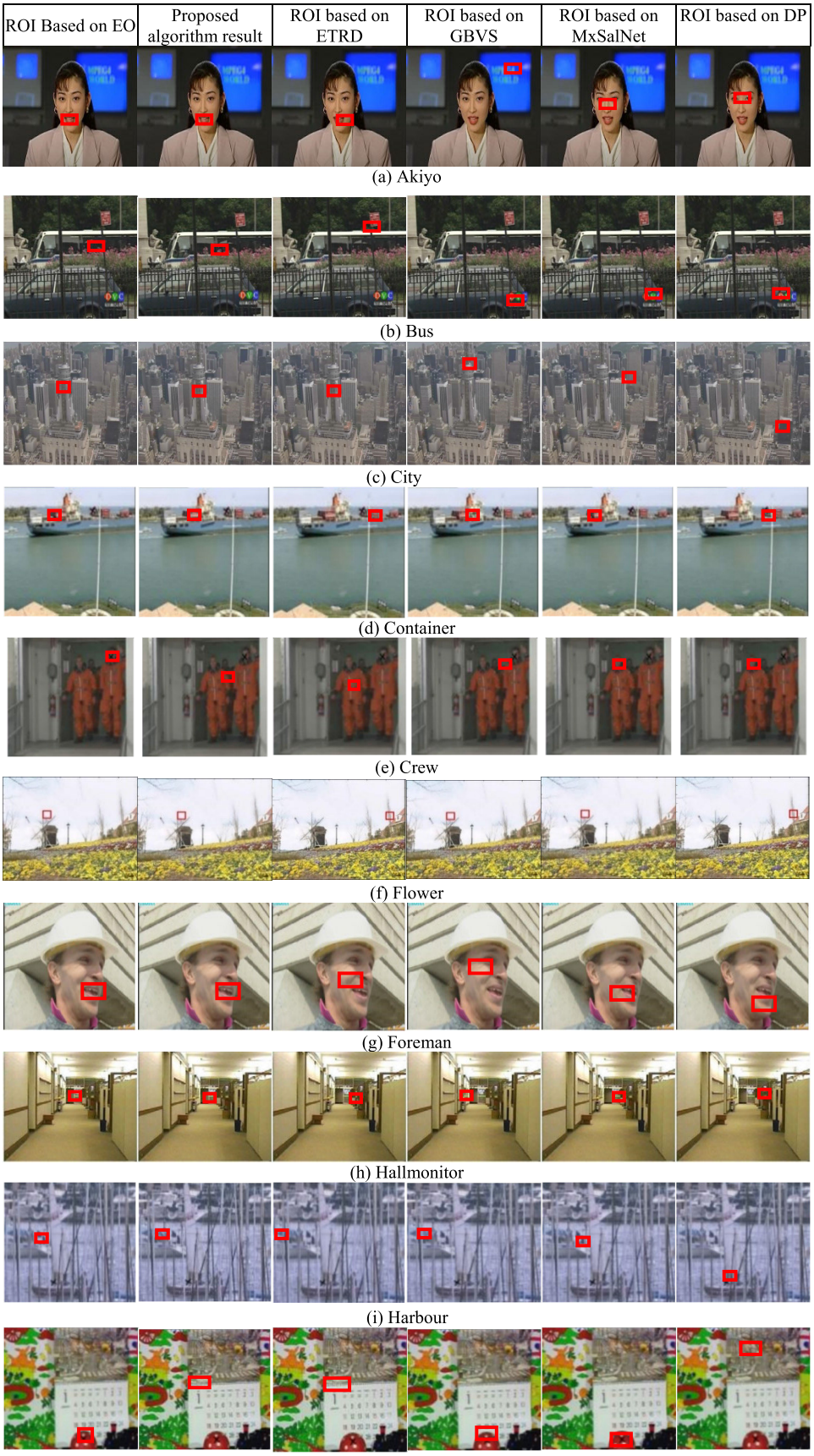


FIGURE 16. Compare ROI of a random frame of videos with respect to proposed algorithm, eye track recorded data, graph based visual saliency, deep learning based visual saliency, depth pixel based visual saliency and expert opinion.

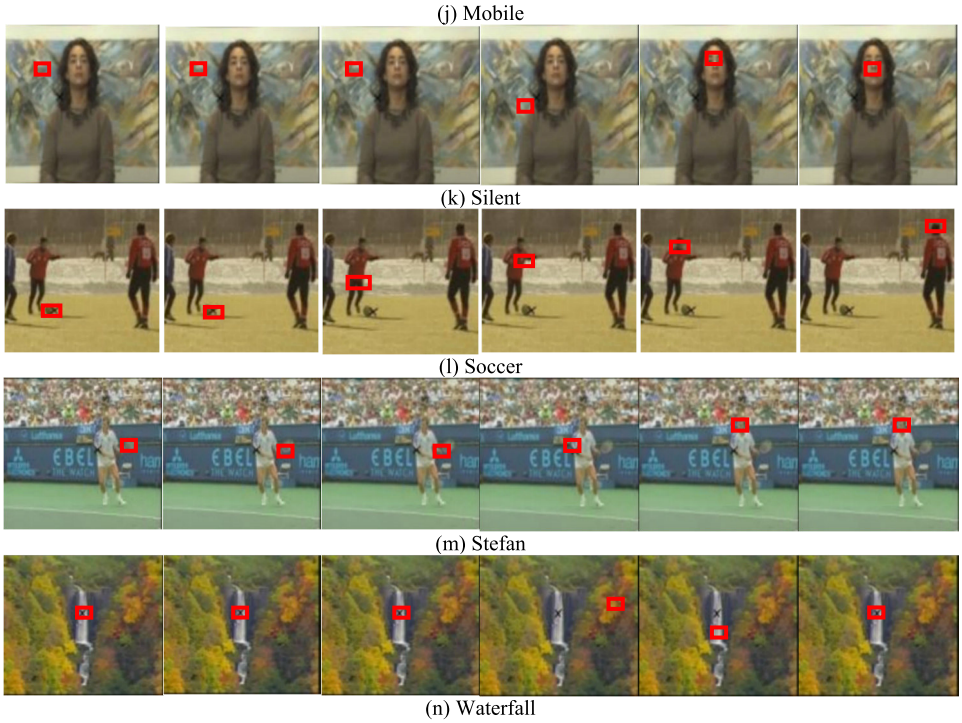


FIGURE 16. *(Continued.)* **Compare ROI of a random frame of videos with respect to proposed algorithm, eye track recorded data, graph based visual saliency, deep learning based visual saliency, depth pixel based visual saliency and expert opinion.**

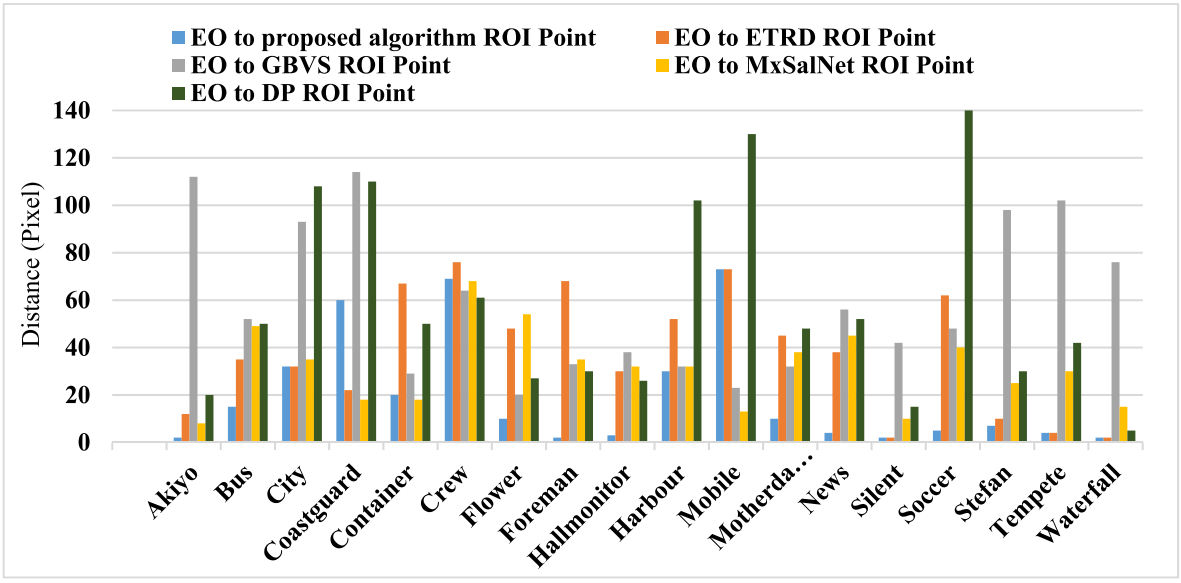


FIGURE 17. **Distance from Experts' Opinion (EO) based ROI to the ETRD, GBVS, MxSalNet, Depth pixel and proposed algorithm based ROI (in pixels).**

as in-focus regions. Moreover, context aware saliency can also be applied that may help to produce compact, appealing, and informative summaries of the videos. Therefore, for compression purposes the following parameters would be considered for the determination of ROI with dominant salience (i) Center sensitivity, (ii) In-focus region, and (iii) Context.

$$S_M = f(S_C, R_I, C_S) \tag{4}$$

where S_M denote the ROI modification parameters, S_C, R_I, and C_s are the centre sensitivity, in-focus region, and context-based saliency respectively.

Only GBVS cannot provide actual ROI of frame and also when both GBVS and ETRD are available, there is no method to reach ultimate single decision for actual ROI. For this reason, a new algorithm is proposed in this paper for determination of ROI based on subjective (ETRD) and objective (GBVS) model which can meet the ROI of human.

Here, distance from the centre of ROI i.e. distance centre to ETRD-ROI, ρ and distance centre to GBVS-ROI τ and tyring concept are used for satisfying centre sensitivity S_C. To consider in focus region R_I, fixation f is considered which is the measurement of when and how much time a participant gaze on a region. Beside this, motion $\exists$ detection is used to understand the context C_S of the video.

The determined point and its surrounding 20% area is considered as ROI for better visualization.

IV. EXPERIMENTAL RESULT

Using this algorithm, we can determine the ROI of an image/video frame which is more likely to the ROI based on EO than ROI based on ETRD and GBVS and those are depicted in **FIGURE 16**. Here, red marked area is the ROI according to the method mentioned in the column heading.

To verify proposed algorithm's output, the distance of ETRD and GBVS based ROI from EO based ROI for each video is presented in **Figure 17**.

For the final result calculation, the deep learning approach MxSalNet, Object oriented approach depth pixel-based saliency (DP), GBVS, ETRD, and the proposed method have been compared to determine the close proximity of the ROIs determined by five different algorithms with the Expert opinion-based benchmark ROI. The outcome reveals that the proposed method outperforms the rest of the techniques in most cases, which is demonstrated in FIGURE 17. If we have available eye track data of any video, proposed algorithm can be used to select more effective ROI and it can be applied in the areas of video compression, computer vision, and image segmentation.

V. CONCLUSION

Region of interest (ROI) can be determined by using both human visual features (Subjective) and mathematical modeling (Objective). If both methods are applied to an image or video frame to determine ROI, technically they should provide the same output. However, in most cases, a clear disparity in result exists. As there is no existing method to determine a single solution when both options are available, in this work, a robust ROI decision algorithm is proposed to determine the ultimate ROI based on the knowledge of subjective and objective information. Experimental results show that for a wide range of video sequences and compared to the existing deep learning based (MxSalNet) and depth pixel (DP) based ROI, the proposed ROI is more consistent to the benchmark ROI, which was previously decided by a group video coding expert. As this work provides more accurate ROI, it can be applied in the areas of video compression to develop more compressed quality video, medical image analysis, image segmentation and such contemporary applications.

REFERENCES

[1] L. Larsson, M. Nyström, R. Andersson, and M. Stridh, "Detection of fixations and smooth pursuit movements in high-speed eye-tracking data," *Biomed. Signal Process. Control*, vol. 18, pp. 145–152, Apr. 2015.

[2] S. Sridharan, R. Bailey, A. McNamara, and C. Grimm, "Subtle gaze manipulation for improved mammography training," in *Proc. Symp. Eye Tracking Res. Appl.*, Mar. 2012, pp. 75–82.

[3] R. S. Kushalnagar, W. S. Lasecki, and J. P. Bigham, "Accessibility evaluation of classroom captions," *ACM Trans. Accessible Comput.*, vol. 5, no. 3, pp. 1–24, Jan. 2014.

[4] K. Gidlöf, A. Wallin, R. Dewhurst, and K. Holmqvist, "Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment," *J. Eye Movement Res.*, vol. 6, no. 1, pp. 1–14, Jan. 2013.

[5] A. Ammar, O. Bouattane, and M. Youssfi, "Automatic cardiac cine MRI segmentation and heart disease classification," *Comput. Med. Imag. Graph.*, vol. 88, Mar. 2021, Art. no. 101864.

[6] R. Hans and H. Kaur, "Hybrid biogeography-based optimization and genetic algorithm for feature selection in mammographic breast density classification," *Int. J. Image Graph.*, vol. 22, no. 3, Feb. 2021, Art. no. 2140007.

[7] P. L. K. Mantos and I. Maglogiannis, "Sensitive patient data hiding using a ROI reversible steganography scheme for DICOM images," *J. Med. Syst.*, vol. 40, no. 6, pp. 1–17, May 2016.

[8] H. Meuel, M. Munderloh, and J. Ostermann, "Low bit rate ROI based video coding for HDTV aerial surveillance video sequences," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2011, pp. 13–20.

[9] P. K. Podder, M. Paul, and M. Murshed, "QMET: A new quality assessment metric for no-reference video coding by using human eye traversal," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Nov. 2016, pp. 1–6.

[10] P. K. Podder, M. Paul, and M. Murshed, "A novel quality metric using spatiotemporal correlational data of human eye maneuver," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2017, pp. 1–8.

[11] P. K. Podder, M. Paul, and M. Murshed, "EMAN: The human visual feature based no-reference subjective quality metric," *IEEE Access*, vol. 7, pp. 46152–46164, 2019.

[12] P. K. Podder, M. Paul, and M. Murshed, "A novel no-reference subjective quality metric for free viewpoint video using human eye movement," in *Proc. Pacific-Rim Symp. Image Video Technol.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 10749, 2018, pp. 237–251.

[13] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2006, pp. 1–8.

[14] M. S. Martin, B. Huard-Nicholls, and A. P. Johnson, "Gaze and pupil size variability predict difficulty-level and safe intersection crosses in a driving simulator," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2021, vol. 65, no. 1, pp. 843–847.

[15] J. Chen, H.-W. Huang, P. Rupp, A. Sinha, C. Ehmke, and G. Traverso, "Closed-loop region of interest enabling high spatial and temporal resolutions in object detection and tracking via wireless camera," *IEEE Access*, vol. 9, pp. 87340–87350, 2021.

[16] X. Zhang, S.-H. Seo, and C. Wang, "A lightweight encryption method for privacy protection in surveillance videos," *IEEE Access*, vol. 6, pp. 18074–18087, 2018.

[17] R. Caldara and S. Miellet, "IMap: A novel method for statistical fixation mapping of eye movement data," *Behav. Res. Methods*, vol. 43, no. 3, pp. 864–878, Sep. 2011.

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

[19] 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.

[20] B. Murugaraj and J. Amudha, "Performance assessment framework for computational models of visual attention," in *Proc. Int. Symp. Intell. Syst. Technol. Appl.*, in Advances in Intelligent Systems and Computing, vol. 683, 2017, pp. 345–355.

[21] M. Paul and M. M. Salehin, "Spatial and motion saliency prediction method using eye tracker data for video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 6, pp. 1856–1867, Jun. 2019.

[22] M. M. Salehin and M. Paul, "A novel framework for video summarization based on smooth pursuit information from eye tracker data," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 692–697.

[23] M. M. Salehin and M. Paul, "Human visual field based saliency prediction method using eye tracker data for video summarization," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2016, pp. 1–6.

[24] A. K. Roy, M. N. Akhtar, M. Mahadevappa, R. Guha, and J. Mukherjee, "A novel technique to develop cognitive models for ambiguous image identification using eye tracker," *IEEE Trans. Affect. Comput.*, vol. 11, no. 1, pp. 63–77, Jan. 2020.
[25] C. Schulze, R. Frister, and F. Shafait, "Eye-tracker based part-image selection for image retrieval," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2013, pp. 4392–4396.
[26] O. Kachan and A. Onuchin, "Topological data analysis of eye movements," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1398–1401.
[27] M. Zhong, X. Zhao, X.-C. Zou, J. Z. Wang, and W. Wang, "Markov chain based computational visual attention model that learns from eye tracking data," *Pattern Recognit. Lett.*, vol. 49, pp. 1–10, Nov. 2014.
[28] K. Kurzhals and D. Weiskopf, "Space-time visual analytics of eye-tracking data for dynamic stimuli," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2129–2138, Dec. 2013.
[29] T. Busjahn, G. Shchekotova, M. Antropova, C. Schulte, B. Sharif, B. Simon, A. Begel, M. Hansen, R. Bednarik, P. Orlov, and P. Ihantola, "Eye tracking in computing education," in *Proc. 10th Annu. Conf. Int. Comput. Educ. Res. (ICER)*, 2014, pp. 3–10.
[30] H. Uwano, M. Nakamura, A. Monden, and K.-I. Matsumoto, "Analyzing individual performance of source code review using reviewers' eye movement," in *Proc. Symp. Eye Tracking Res. Appl.*, 2006, pp. 133–140.
[31] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2014, pp. 5040–5043.
[32] H. Hadizadeh, M. J. Enriquez, and I. V. Bajic, "Eye-tracking database for a set of standard video sequences," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 898–903, Feb. 2012.
[33] L. Jia, X. Zhong, and Y. Tu, "No reference video quality assessment model based on eye tracking datas," in *Proc. 2nd Int. Conf. Inf., Electron. Comput.*, vol. 59, Mar. 2014, pp. 97–100.
[34] S. F. Dodge and L. J. Karam, "Visual saliency prediction using a mixture of deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4080–4090, Aug. 2018.
[35] K. Zhou, A. Paiement, and M. Mirmehdi, "Detecting humans in RGB-D data with CNNs," in *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2017, pp. 306–309.
[36] L. Zhang and H. Wu, "Cosaliency detection and region-of-interest extraction via manifold ranking and MRF in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2021.
[37] S. Ma, W. Zhu, and L. Zhang, "Region of interest extraction based on unsupervised cross-domain adaptation for remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2021, pp. 1169–1172.
[38] G. Sun, J. Zhang, K. Zheng, and X. Fu, "Eye tracking and ROI detection within a computer screen using a monocular camera," *J. Web Eng.*, vol. 19, nos. 7–8, pp. 1117–1146, Dec. 2020.
[39] K. He, C. Yang, V. Stankovic, and L. Stankovic, "Graph-based clustering for identifying region of interest in eye tracker data analysis," in *Proc. IEEE 19th Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2017, pp. 1–6.
[40] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 226–231.
[41] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, 2000.
[42] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2012.
[43] O. Špakov and D. Miniotas, "Application of clustering algorithms in eye gaze visualizations," *Inf. Technol. Control*, vol. 36, no. 2, pp. 213–216, Jun. 2007.
[44] A. Santella and D. DeCarlo, "Robust clustering of eye movement recordings for quantification of visual interest," in *Proc. Eye Track. Res. Appl. Symp.*, 2004, pp. 27–34.
[45] H. Öğmen and B. G. Breitmeyer, *The First Half Second: The Microgenesis and Temporal Dynamics of Unconscious and Conscious Visual Processes*. Cambridge, MA, USA: MIT Press, 2006, p. 410.
[46] A. Bruno, F. Gugliuzza, E. Ardizzone, C. C. Giunta, and R. Pirrone, "Image content enhancement through salient regions segmentation for people with color vision deficiencies," *iPerception*, vol. 10, no. 3, May 2019, Art. no. 2041669519841073.
[47] *Tobii Eye Tracker Manual, Tobii Studio 2.2*. Accessed: Sep. 2010. [Online]. Available: https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf
[48] J. Chilukamari, S. Kannangara, and G. Maxwell, "A low complexity visual saliency model based on in-focus regions and centre sensitivity," in *Proc. IEEE 4th Int. Conf. Consum. Electron. Berlin (ICCE-Berlin)*, Sep. 2014, pp. 411–414.
[49] P. K. Podder, M. Paul, T. Debnath, and M. Murshed, "An analysis of human engagement behaviour using descriptors from human feedback, eye tracking, and saliency modelling," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Nov. 2015, pp. 1–8.

MANORANJAN PAUL (Senior Member, IEEE) received the Ph.D. degree from Monash University, in 2005. He worked as a Research Fellow with the University of New South Wales, Monash University, and Nanyang Technological University. He is currently working as a Professor of computer science, the Director of the Computer Vision Laboratory, and the Head of the Machine Vision and Digital Health Research Group, Charles Sturt University. He has published more than 220 fully refereed international publications, including more than 90 journals and supervised 17 Ph.D. in completion, including six as a principal supervisor. He has obtained $15 million competitive external grant money, including Australian Research Council (ARC) Discovery Project (DP19 and DP13), Soil CRC, NSW Government, Wine Australia, Western Australia DPIRD, and NSW DPI grants. His research interests include image/video coding, EEG signal analysis, and computer vision. He is a Senior Member of the Australian Computer Society (ACS). He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and *EURASIP Journal on Advances in Signal Processing*. He was a Keynote Speaker in IEEE DICTA-17, WoWMoM-14 Workshop, DICTA-13, and ICCIT-10. He is the ICT Researcher of the Year 2017 by ACS.

PALLAB KANTI PODDER (Member, IEEE) received the Ph.D. degree from the CSU Machine Learning Research Unit, Charles Sturt University, Bathurst, NSW, Australia, in December 2017. Then, he joined as an Assistant Professor with the Department of Information and Communication Engineering, Pabna University of Science & Technology (PUST), which is one of the leading science and technology universities in Bangladesh. He has acted as an Advisor IEEE (PUST Branch), and the Director of the Software Engineering Laboratory, PUST. Then, he came back to Australia as a Postdoctoral Research Fellow and casual teaching academic at Charles Sturt University. He has published more than 40 journal articles and international conference proceedings in the areas of image processing, video compression, and video quality assessment.

MD. RIAD HASSAN (Student Member, IEEE) received the B.Sc. (Engineering) degree (Hons.) in information and communication engineering from the Pabna University of Science and Technology (PUST), Pabna, Bangladesh. He is currently pursuing the master's degree with the Bangladesh University of Engineering and Technology (BUET). He is also working as a Lecturer with the Green University of Bangladesh. Beside this, he is conducting research in collaboration with Charles Sturt University, Australia. His research interests include image processing, video compression, video quality assessment, and medical imaging. He is a member of the Bangladesh Computer Society. He has been awarded research grand, worth $10,000 from the ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of Bangladesh, and also received the University Merit Scholarship in consecutive four years during his undergrad studies. Recently, he has received Academic Excellence Award from the PUST in 2022.