IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 December 2022, accepted 8 February 2023, date of publication 13 February 2023, date of current version 17 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244191

RESEARCH ARTICLE

Saliency Prediction in Uncategorized Videos Based on Audio-Visual Correlation

MARYAM QAMAR[1,2,3], SULEMAN QAMAR[4], MUHAMMAD MUNEEB[5], SUNG-HO BAE[6], (Member, IEEE), AND ANIS RAHMAN[1]

[1]Department of Computing, National University of Science and Technology, Islamabad 44000, Pakistan
[2]Department of Computer Science and Information Technology, University of Azad Jammu and Kashmir, Muzaffarabad 13100, Pakistan
[3]Department of Artificial Intelligence, Kyung Hee University, Seoul 17104, South Korea
[4]Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Islamabad 45600, Pakistan
[5]Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates
[6]Department of Computer Science and Engineering, Kyung Hee University, Seoul 17104, South Korea

Corresponding author: Suleman Qamar (m.sulemanqamar@gmail.com)

This work was supported by the Department of Mathematics, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates.

ABSTRACT Substantial research has been done in saliency modeling to make intelligent machines that can perceive and interpret their surroundings and focus only on the salient regions in a visual scene. But existing spatio–temporal saliency models either treat videos as merely image sequences excluding any audio information or are unable to cope with inherently varying content. Based on the hypothesis that an audiovisual saliency model will perform better than traditional spatio–temporal saliency models, this work aims to provide a generic preliminary audio/video saliency model. This is achieved by augmenting visual saliency map with an audio saliency map computed by synchronizing low-level audio and visual features. The proposed model was evaluated using different criteria against eye fixations data for a publicly available video dataset DIEM. The evaluation results show that the model outperforms two state-of-the-art visual spatio–temporal saliency models. Thus, supporting our hypothesis that an audiovisual model performs better in comparison to a visual model for natural uncategorized videos.

INDEX TERMS Saliency, audiovisual, uncategorized videos, spatio–temporal.

I. INTRODUCTION

Humans are able to quickly identify and analyze the most intriguing elements of a complex visual environment because of the perceptual and cognitive mechanisms termed as visual attention. Rapid eye movements, or saccades, are constantly used to quickly scan new things of interest and thus analyze the scene. These mechanisms aid in the prioritization and filtering of stimuli as they go from the initial stages of visual processing to later stages, which are capable of higher-level cognitive processing. Effective scene scanning is made possible by the ability of humans to recognize objects' saliency in a visual scene [1]. Simulating the visual attention mechanism in machines is termed as visual saliency prediction. Automatic saliency prediction has applications in several research and practical fields including computer vision, robotics, healthcare, and multimedia [2], [3], [4], [5].

Many saliency computing algorithms designed for images [6], [7], [8] are around that use visual cues, for example color, intensity, orientation etc., other models [9], [10], [11] also take social cues like face into account and are found to give more accurate eye movement predictions. spatio-temporal saliency models for image sequences [12], [13], [14] usually incorporate temporal cues like motion but ignore the effect of audio stimuli, an integral part of video content, on human gaze, and hence such models can be classified as unimodal models [15] where only visual stimuli is used.

Interestingly the effect of audio stimuli is proved to be relevant to human eye movements, for instance in [16] the authors find eye movements to be spatially biased towards the source of audio by performing an eye tracking experiment on images with spatially localized sound sources in three conditions: auditory (A), visual (V) and audio-visual (AV). Moreover, another study [17] analyzed the effects of different type of sounds on human gaze involving an experiment with 36 participants and thirteen sound classes under audio-visual

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

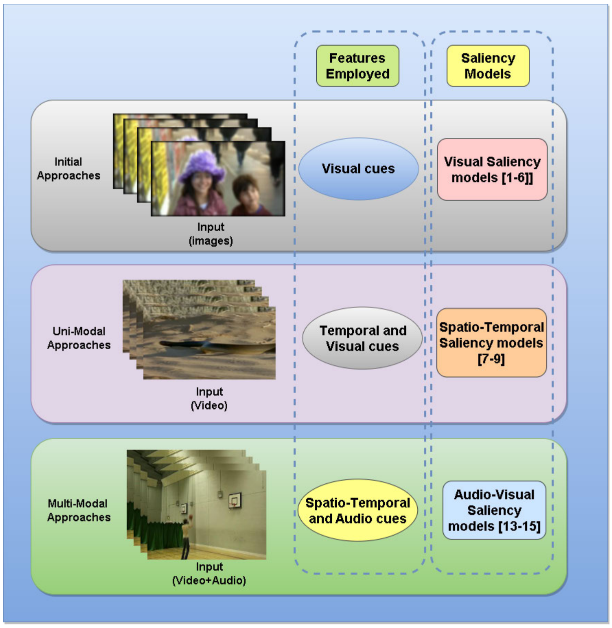

FIGURE 1. Historical Perspective: Traditional Approaches employing visual cues to Uni-model approaches for video saliency where temporal cues are added to multi-modal methods employing audio along with Spatio-temporal cues.

and visual conditions. The sound classes are further clustered into on-screen with one sound source, on-screen with more than one sound source and off-screen sound source. Kullback-Leibler divergence is used to evaluate eye positions and fixation durations between the two conditions. The results show that human speech, singer(s) and human noise (on-screen sound source clusters) highly affect gaze and, more importantly, linked audio-visual stimuli has a greater effect than unsynchronized audio-visual events. Though a lot of research has been done in the general field of unimodal models for both images and videos [18], [19], no remarkable contribution is found in the area of bimodal saliency modeling. Of more consequence is the lack of a model for computation of audiovisual saliency in complex video sequences. Existing literature in the area of audio-video saliency modeling is scarce and often aims at a specific class of videos [20], [21], [22]. Motivated by the lack, an extension of a typical visual saliency model into an audio-visual model to predict salient regions in complex videos with different sound classes is required.

The focus of this work is to propose a generic audio-visual saliency model for complex video sequences. The work differs from previous research [20], [21], [22] in that it does not restrict input videos to be from a certain category. To accomplish that an audio source localization method was used to relate an audio signal with an object in the video frames in a rank correlation space. The proposed model was evaluated against eye fixations ground truth from DIEM dataset.

A. NOVEL CONTRIBUTION

The original contribution of this study is as follows:

1) An audio-visual saliency model for complex scenes that, unlike existing literature, does not restrain videos to any specific category and experimental evaluation of same for the purpose of analyzing the effect of audio stimuli.
2) Presented and analyzed the results of preliminary experimental evaluation on a publicly available dataset to examine how our proposed saliency model compares to two state-of-the-art saliency models.

B. ORGANIZATION

The remainder of the paper is organized as follows: Section I narrated the background knowledge of saliency modeling and novel contribution of this work towards it. Section II gives a detailed review of state-of-the-art literature pertaining to the problem under consideration and their drawbacks while Section III describes the proposed solution step by step. Section IV provides implementation details of the proposed solution and outlines the properties of video sequences in the dataset used for experimentation, this section also explains the saliency evaluation metrics used for evaluation of proposed model. Section V presents the performance results of the proposed solution on the dataset described in the previous section followed by a discussion in Section VI. Section VII summarizes our main findings and concludes by indicating future research directions.

II. RELATED WORK

Unimodal saliency models use only one type of sensory stimulus as input traditionally some visual cues including color, intensity and orientation features [8], [23], [24]. Other biologically-inspired models [14], [25] exploit spatial contrast and motion, and simulate interactions between neurons using excitation and inhibition mechanisms. While some spatio-temporal models [26], [27] propagate spatial/temporal saliency making use of multiscale color and motion histograms as features. In [26] pixel-level spatio-temporal saliency is computed from spatial and temporal saliencies via interaction and selection driven from superpixel-level saliency. Reference [27] propagates temporal saliency forward and backward via inter-frame similarity matrices and graph-based motion saliency, whereas spatial saliency is propagated over a frame using temporal saliency and intra-frame similarity matrices. In most of these models, conspicuity maps are constructed by a variety of approaches using different visual features that are later integrated together via a number of fusion schemes to get a final saliency map.

Based on the fact that eyes are the most important sensory organs providing much of the information around humans, many state-of-the-art visual models [26], [27] aim at saliency computation for complex dynamic scenes. But such unimodal models tend to overlook other influential social cues like faces in social interaction scenes, and hence exhibiting lower predictability [28], [29]. Moreover, social scenes involve a lot more sensory signals influencing eye movements spatially such as auditory information including voice tone, music, etc [30]. In addition, different kinds of sounds affect eye fixations differently [16], [17]. Thus, soliciting the need of a bimodal saliency model that incorporates both visual and audio information channels.

Rapantzikos et al. [31] proposed an audio-visual saliency model for movie summarization. The visual saliency map is constructed using traditional features i.e. intensity, color and motion and simulating feature competition as energy minimization via gradient descent. Which is then thresholded and average saliency per frame is computed to construct a $1D$ visual saliency curve. While maximum average Teager energy, mean instant amplitude and mean instant frequency are extracted as audio features by applying Teager-Kaiser Energy Operator and Energy Separation Algorithm on the audio signal. The audio feature vector is normalized to the range $[0, 1]$ and a weighted fusion is applied to achieve audio saliency curve. Final audio-visual saliency curve is a weighted linear combination of audio and visual saliency curves. Local maxima feature of audio-visual saliency curve is used for key-frame selection. The experiments are conducted on movie database of A.U.T.H but no comparison and evaluation is given.

Coutrot and Guyader [32] proposed an audiovisual saliency model for natural conversation scenes; a linear combination of low level saliency, face map and center bias. Low level saliency map is constructed via Marat's spatio-temporal saliency model [14]. While for face map construction a speaker diarization algorithm is proposed which uses motion activity of faces and 26 Mel Frequency Cepstral Coefficients (MFCCs) as visual and audio features respectively. Center bias is a time-independent 2D Gaussian function centered at screen center. The three maps are linearly combined into final audiovisual saliency map using expectation maximization to decide the weight for each. Normalized Scanpath Saliency (NSS) score showed that proposed model performs better than same model without speaking and mute face differentiation but target video dataset belongs to a limited category; conversation scenes only.

Ould-Sidaty et al. [22] proposed an audiovisual saliency model for teleconferencing and conversational videos. Three best performing models on target database i.e. Itti et al. [33], Harel et al. [34], and Tavakoli et al. [35] are selected as spatial model. Acoustic energy is computed per frame and block matching algorithm is used to construct visual features from the face stream of video for audio map. Then peak matching is used for audio-visual synchronization. Five fusion schemes are used to get final map and Global Non-Linear Normalization followed by Normalization performed best. Experiments performed on XLIMedia database created by authors showed that proposed model performed better than all three spatial models. Limitation of this work is that it only targets conferencing and conversational videos.

In [36] authors detect the spatial and temporal saliency maps from the visual modality then they use cross-modal kernel canonical correlation analysis to compute the audio saliency map from both modalities by localizing the moving-sounding objects. They also propose a two-stage adaptive audiovisual saliency fusion method to integrate the spatial, temporal and audio saliency maps to audio-visual saliency map. This work has collected an audio-visual dataset for video saliency prediction, the main difference between this and our work is our evaluation dataset does not restrict the number of sound sources and also includes background sounds which makes the task more challenging while their dataset only have videos with one sound source.

A brief historical perspective of saliency models is shown in the figure 3. All in all, one of the major drawback of visual models is that such models treat videos as a mute sequence of images ignoring the influence of audio stimuli resulting in inaccurate predictions often in such cases where sound guides eye movement. Furthermore, there is a knowledge gap due to the absence of any audiovisual model for complex dynamic scenes; that is, many of the state-of-the-art audiovisual models restrict the dataset used to only one specific category of audio, for instance, conversational videos, thus limiting the models' performance when dealing with videos containing different sound classes.

III. PROPOSED SOLUTION

This section explains the proposed solution for audio-visual saliency computation for videos. The framework consists of five major modules, the first one, feature extractor, takes audio and visual stimuli and outputs audio energy descriptors and object motion descriptors per frame by processing the stimuli through separate channels explained in next sections. Next module computes audio saliency map from the audio and motion descriptors produced previously while visual saliency map computation and motion map computation can work in parallel computing visual map from low level features like intensity, color, orientation etc. and motion map from color-coded optical flow previously computed while processing visual stimuli in first step of the solution. Then all the maps are normalized and combined in a unified audiovisual saliency map. Figure 2 outlines the proposed framework.

A. FEATURE EXTRACTION

This is the first step of proposed methodology where a number of visual and acoustic features were extracted from a given input video by using a modified version of the method given by [37]. The step comprises of two branches for feature extraction, audio features and visual features respectively. A detailed workflow is shown in figure 3

1) AUDIO FEATURE EXTRACTION

Audio energy descriptor $a(t)$ of an audio signal featuring the changing patterns of an audio signal strength in the same temporal resolution as the corresponding video frames was obtained. In detail, the audio signal was first segmented into frames according to the frame rate of video such that each audio frame corresponds to a video frame. Using STFT (short-term Fourier Transform), this framed signal was transformed into time-frequency domain to get a spectrogram of the signal at each frame. The audio energy of the input signal $a(t)$ was computed by the integration of the resultant spectrogram at any given frame over all frequencies using,

$$a(t) = \int_0^\infty \int_0^T f(t')W(t'-t)e^{-j2\pi ft'}dt'df$$

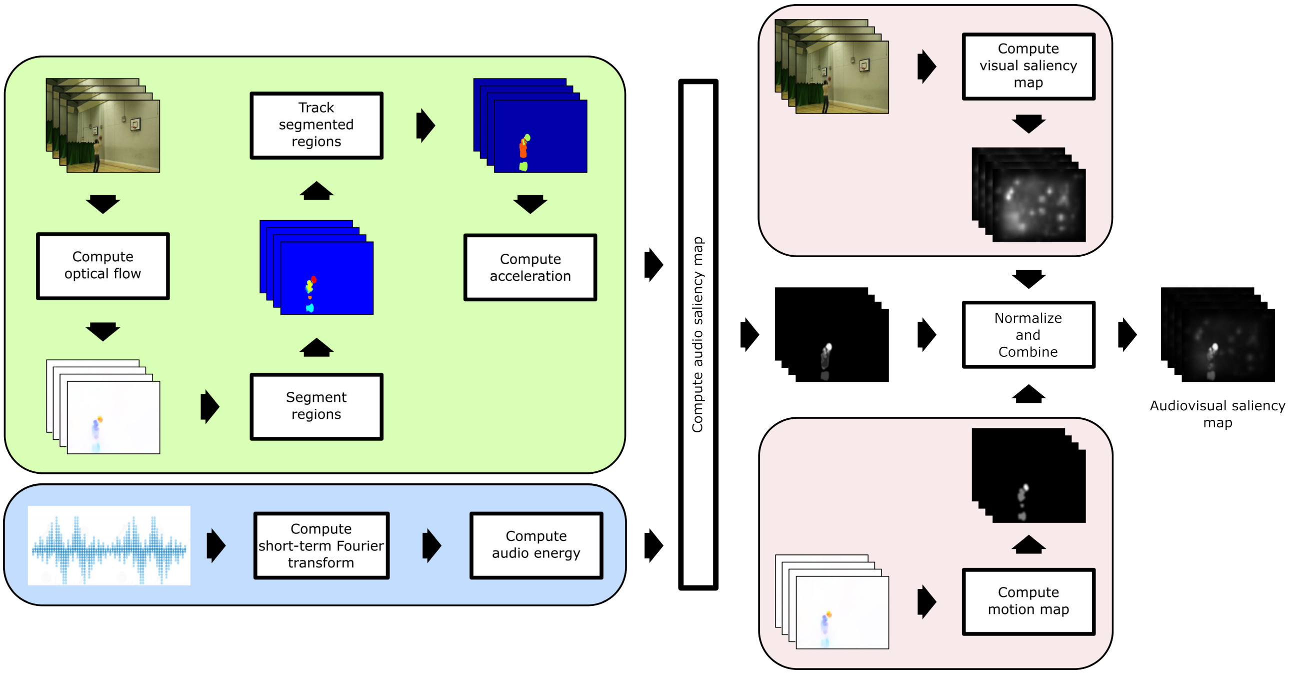

FIGURE 2. Architecture of proposed solution: top left - video instance segmentation and visual features computation, bottom left - audio features computations from audio signal, top right - bottom up visual saliency map computation, bottom right - motion map computation from optical flow.

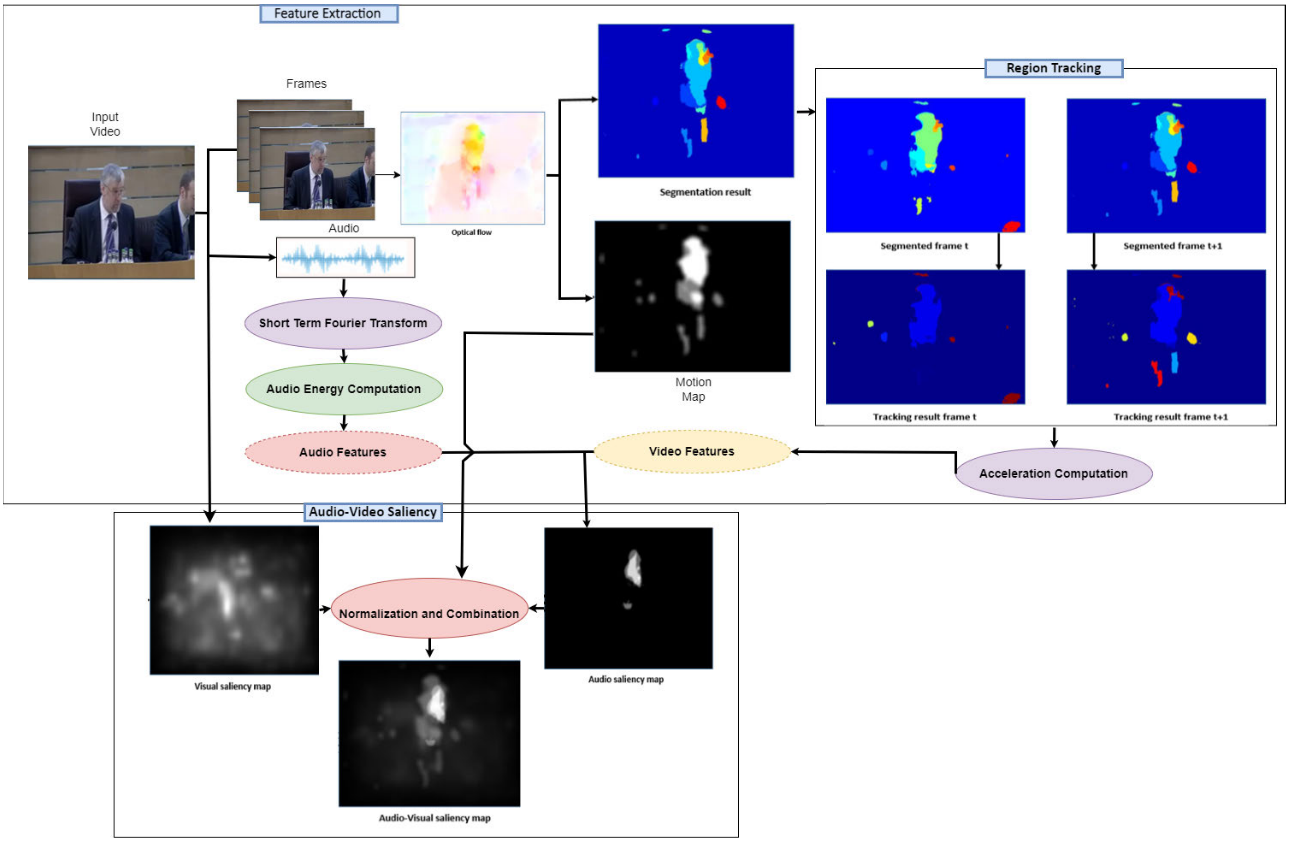

FIGURE 3. Proposed solution workflow: Feature Extraction - video frames are used to calculate optical flow, which is used for both video features and motion maps computation, audio features i.e. audio energy is calculated via STFT. Audio saliency maps are generated via correlation of these features. Audio-Visual Saliency - Audio-visual Saliency is then acquired by combining the normalized audio, visual and motion maps.

where windowing function $W(t)$ is defined such that the neighboring windows overlap by 50%. This audio energy descriptor $a(t)$ was then filtered using a $1D$ Gaussian kernel.

2) VISUAL FEATURE EXTRACTION

Based on the assumption that in a video any moving object is a prime candidate to be an audio signal source, proposed solution aims at performing unsupervised video instance segmentation to get candidate objects for correlating with audio stimuli later. This task includes both segmentation and tracking of objects through all video frames but it is a relatively new and non-trivial computer vision task [38], [39]. Techniques for solving this problem are just emerging now and have not been tested rigorously. On the other hand, video object segmentation using optical flow has been put to use many times and proved to perform well [38]. The success of optical flow based methods motivated us to devise a method using it for the purpose of video instance segmentation. Also, sparse optical flow only calculates some edges and corners of objects in order to avoid aperture problem while we need the tracking of whole-body mass of moving object because it is used both in segmentation set up and motion map computation. Segmented objects obtained using optical flow were then tracked along all frames via color histograms of the regions in HSV color-space and are then used to calculate acceleration per frame of all the moving objects in a given input video, designated as motion descriptor. The process is described in detail as follows:

1) **Optical Flow Computation.** The method proposed by [40] was used to compute dense optical flow and corresponding color-coded optical flow images for all frames of a given input video. This method uses apparent motion of each pixel to compute forward and backward optical flows where the former depicts the motion of pixels of frame t with reference to frame $t+1$ and the latter is the motion of pixels of frame t with respect to frame $t-1$. These resulting optical flows were then averaged to get the mean optical flow per frame, which was used as a basis for performing video instance segmentation to be then used in calculation of visual features employed in computing audio saliency map through the correlation of audio and visual features.
2) **Frame Segmentation.** The color coded mean optical flow image corresponding to each frame is used as input for the segmentation step. Meanshift, a nonparametric clustering algorithm, is applied for segmentation of each input image in LUV color space. The over-segmented result of the Meanshift segmentation step is followed by a simple region merging technique based on DeltaE, a color difference score, to merge the closely similar regions. Regions smaller than 200 pixels are filtered then to remove noisy and insignificant regions.
3) **Region Tracking.** Once individual frames are segmented, a number of tracks are initialized using regions' location and appearance features in the first frame. Then all regions in each new segmented input frame are either assigned to an existing track or initialized to a new track based on its location and appearance similarities. The location similarity d_E is computed by Euclidean distance between the centroid of a new region C_n and that of an existing track C_e using,

$$d_E = \sqrt{(C_n(x,y) - C_e(x,y))^2}$$

The similarity measure outputs a list of candidate tracks similar to the region under consideration for assignment within a specified search radius r. For appearance similarity AS LUV histograms of these existing candidate tracks H_e are compared with the new region's histogram H_n using cosine value of angle $cos\theta$ as,

$$cos\theta = \frac{H_n \cdot H_e}{||H_n||||H_e||}$$

The region is assigned to the track whose cosine value of angle is maximum of all candidate tracks and greater than a specified threshold. The track is then updated by replacing the centroid with the centroid of newly assigned region and histogram with the mean of existing histogram and new region's histogram. Otherwise if this maximum value is less than the specified threshold the region is used to initialize a new track.
4) **Calculate Acceleration.** In this step segmented and tracked objects' acceleration is computed using the aforementioned motion descriptors. Average of forward and backward optical flow gives the acceleration at each pixel (x, y, t) in a frame given by equation:

$$g(x,y,t) = F^{+}(x,y,t) + F^{-}(x,y,t)$$

where x and y are spatial coordinates, t is frame number and F^{+} and F^{-} indicate forward and backward optical flow.
The acceleration of regions ST_i^t where i is region index per frame t is computed as the average acceleration of all pixels belonging to that region as:

$$m_i(t) = \frac{1}{|ST_i^t|} \sum_{(x,y)\varepsilon ST_i^t} ||g(x,y,t)||$$

The acceleration vector is filtered with Gaussian kernel to remove visual noise to get motion descriptor of objects in a given input video.

B. AUDIO SALIENCY MAP COMPUTATION

For the audio saliency map computation, the method proposed in [37] for audio-video correlation is used. Humans are incredibly good at matching the sounds they hear with the accompanying visual perception, which allows them to locate and separate various sounding objects in a scene. The task of sound source localization in visual scenes is to mimic humans' such ability, where the goal is to identify regions of a visual that correlate strongly with the audio signals [41]. In this work, the correlation between audio and motion descriptors computed in previous steps is used to localize the source of sound signal in given video frames to indicate audio saliency. Winner-Take-All (WTA) hash a subfamily of hashing functions [42] controlled by the number of permutations N and window size S is used to transform both feature vectors in rank correlation space. Once both descriptors are in the common rank correlation space Hamming distance is used to relate the audio signal with the object having maximum synchronization.

C. VISUAL SALIENCY MAP COMPUTATION

A basic saliency map which takes only visual input for saliency computation is used here proposed in [34]. The model approaches the problem of saliency computation by defining Markov chains over feature maps, extracted for features of intensity, color, orientation, flicker and motion, and treat equilibrium locations as saliency values. In detail each value of the feature map(s) is considered to be a node and the connectivity between them holding weights determined by their dissimilarity, defining Markov chain on this graph, the equilibrium distribution of this chain computed by repeated multiplication of Markov matrix with an initially uniform vector accumulates mass at highly dissimilar nodes giving activation maps. Similar mass concentration process is applied on activation maps and output is summed into the final visual saliency map.

D. MOTION MAP COMPUTATION

Motion map indicates the regions of high motion in a given video frame computed with the mean optical flow for each frame as described in Section III-A2. Adaptive thresholding proposed in [43] is applied on the color-coded mean optical flow frames to discard any inconsequential low motion. The method works by setting any pixel I_p to zero if its brightness is T percent lower than average brightness of the surrounding pixels, otherwise setting it to one.

$$I_p = \begin{cases} 0 & \text{if } I_p < T \cdot I_{avg} \\ 1 & \text{otherwise} \end{cases}$$

E. NORMALIZATION AND COMBINATION

In this final step of the proposed model, the three computed maps: a) visual saliency map, b) audio saliency map, and c) motion map are normalized before combining them together in a final audiovisual saliency map. Here the visual saliency map is a sum of normalized activation maps computed, as explained in Section III-C, using mass concentration algorithm that works like activation map construction algorithm. The other two maps are normalized to a specified range $[0-1]$ using simple linear transformations. The resulting normalized maps are then linearly combined to get the final audiovisual saliency map.

IV. MATERIALS AND METHODS

A. IMPLEMENTATION DETAILS

The proposed solution is implemented in MATLAB 2014b and Windows 10 on a 64-bit architecture machine with Intel i5 processor. The same hardware and software setup is used for evaluation purposes. The parameters used for the proposed solution are given in Table 1.

TABLE 1. Parameters used for different steps of the proposed solution.

	Parameter	Value
Region Tracking	Search Radius (r)	100
2*Audio-Video Corr.	No. of Permutations (N)	2000
	Window Size (S)	5
Motion Map Comp.	Threshold % (T)	10

B. DATASET

Unfortunately, most video saliency datasets and their ground truth exclude audio cues, making publicly available standard audio visual saliency prediction datasets scarce. One available dataset is DIEM (Dynamic Images and Eye Movements) dataset [44], which initially, comprised 26 videos of different genres, most with both audio and video data, converted to 30 frames per second MPEG-4 files. Eye fixation data is collected via binocular eye tracking experiment with 17 male and 25 female (forty-two in total) participants with ages ranging between 18 and 36 years with normal/corrected-to-normal vision. The dataset later extended to 85 videos with eye fixation data recorded for more or less 250 participants. Alongside the dataset, DIEM project also provides a tool, referred to as CARPE (Computational and Algorithmic Representation and Processing of Eye movements), to visualize eye fixation data.

In this work, for evaluation 25 video sequences were randomly selected from DIEM dataset. The video sequences are listed in Table 2 along with some properties.

C. EVALUATION METRICS

The proposed solution was evaluated using the listed four saliency evaluation measures.

1) **Area under the curve** (AUC)**.** Reference [45] is a location-based metric, where the number of fixation pixels is counted, as well as, the same number of pixels are randomly extracted from the saliency map. The true positives (TP) and the false positives (FP) are then calculated for different threshold values treating saliency pixels as a classifier. The resulting values are then used to plot a ROC curve and compute the AUC–the ideal score being 1.0 and a value of 0.5 indicating random classification.
2) **Kullback-Leibler divergence** (D_{KL})**.** is a distribution-based measure of dissimilarity given by the equation,

$$D_{KL} = \sum_i M_f(i) ln \left(\frac{M_f(i)}{M_s(i)} \right)$$

it estimates the loss of information when saliency map M_s is used to approximate fixation map M_f–both considered as probability distributions.

The ideal D_{KL} score is zero, meaning the saliency and fixation maps are exactly same, otherwise, higher the score poorer the saliency model.

TABLE 2. Summary of properties of video sequences selected from DIEM dataset. In Audio source column On-screen(+)/Off-screen(−): H = human, N = non-human, M = music and A = applause. Properties in order are: Single object(−)/Multiple objects(+) (f_1), Camera motion (f_2), Abrupt scene change (f_3), Interaction (f_4), Occlusion (f_5), Deformation (f_6), Crowd (f_7), Clutter (f_8), and Motion blur (f_9). In columns f_2 to f_9 (+) indicates presence and (−) indicates absence of the particular property.

No	Video Sequence	Scene Type	Audio Source	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
1	50_people_brooklyn_1280x720	Other	$H^{+/-}M^{-}$	+	+	+	-	-	-	+	-	+
2	advert_bbc4_bees_1024x576	Advertisement	$M^{-}N^{+}$	+	-	-	-	-	-	-	-	-
3	advert_bbc4_library_1024x576	Advertisement	M^{-}	+	-	-	-	-	-	-	+	-
4	advert_bravia_paint_1280x720	Advertisement	$M^{-}N^{+}$	+	-	+	-	-	+	-	-	-
5	arctic_bears_1066x710	Documentary	$H^{-}M^{-}N^{+}$	+	-	-	-	-	-	-	-	-
6	basketball_of_sorts_960x720	Sports	$M^{-}N^{+}$	+	-	-	+	+	-	-	-	-
7	BBC_wildlife_special_tiger_1276x720	Documentary	$H^{-}M^{-}N^{+}$	-	-	-	-	+	-	-	-	-
8	DIY_SOS_1280x712	Other	H^{+}	-	-	-	-	-	-	-	-	-
9	documentary_adrenaline_rush_1280x720	Documentary	$H^{-}M^{-}$	+	+	-	-	+	-	-	-	-
10	documentary_coral_reef_adventure_1280x720	Documentary	$H^{-}M^{-}N^{+}$	+	+	+	+	+	-	-	-	-
11	game_trailer_lego_indiana_jones_1280x720	Computer Game	$H^{-}M^{-}N^{+}$	+	-	+	+	+	+	+	-	-
12	hairy_bikers_cabbage_1280x712	Other	H^{+}	+	-	-	+	-	-	-	-	-
13	harry_potter_6_trailer_1280x544	Movie	$H^{+}M^{-}N^{+}$	+	-	+	+	+	+	-	-	-
14	home_movie_Charlie_bit_my_finger_again_960x720	Movie	H^{+}	+	+	-	+	-	-	-	-	-
15	hummingbirds_closeups_960x720	Documentary	$H^{-}N^{+}$	+	-	-	+	-	+	-	-	-
16	music_trailer_nine_inch_nails_1280x720	Crowd	$M^{+/-}$	+	-	-	+	+	-	-	-	-
17	news_bee_parasites_768x576	News	$H^{+/-}$	+	-	-	+	+	-	-	-	-
18	news_sherry_drinking_mice_768x576	News	H^{-}	+	-	+	+	+	-	-	-	-
19	news_us_election_debate_1080x600	News	$A^{-}H^{+}$	+	-	-	+	+	-	-	-	-
20	one_show_1280x712	Other	H^{+}	-	+	-	-	-	-	-	-	-
21	pingpong_angle_shot_960x720	Sports	N^{+}	+	-	-	+	-	-	-	-	-
22	planet_earth_jungles_monkeys_1280x704	Documentary	$H^{-}N^{+}$	+	-	-	-	+	-	-	-	-
23	scottish_parliament_1152x864	Other	$H^{+/-}$	+	-	-	-	-	-	+	-	-
24	sport_football_best_goals_976x720	Sports	$A^{-}M^{-}$	+	+	-	+	+	-	-	-	-
25	stewart_lee_1280x712	Other	H^{+}	+	-	-	+	+	-	+	-	-

3) **Normalized Scanpath Saliency** (NSS)**.** Reference [46] is computed using,

$$NSS = \frac{1}{N}\sum_{i}\frac{M_s(i) - \mu_{M_s}}{\sigma_{M_s}}$$

where saliency map M_s is normalized to zero mean and unit standard deviation, which is then averaged for N fixations.
Zero score means a chance prediction whereas a high score indicates high predictability of the saliency model.

4) **Linear Correlation Coefficient** (CC) is another distribution-based metric computed using the equation,

$$CC = \frac{cov(M_s, M_f)}{\sigma_{M_s}\sigma_{M_f}}$$

Its output ranges between -1 and $+1$, the closest is the score to $+1$, the better is predictability of the saliency model.

Performance metrics are summarized in table 3.

TABLE 3. Performance metrics summary.

Name	Abbreviation	Formula
Area under the curve	AUC	-
Kullback-Leibler divergence	D_{KL}	$\sum_{i} M_f(i)ln\left(\frac{M_f(i)}{M_s(i)}\right)$ (1)
Normalized Scanpath Saliency	NSS	$\frac{1}{N}\sum_{i}\frac{M_s(i)-\mu_{M_s}}{\sigma_{M_s}}$ (2)
Linear Correlation Coefficient	CC	$\frac{cov(M_s, M_f)}{\sigma_{M_s}\sigma_{M_f}}$ (3)

D. COMPARISON METHODS

Based on our literature review, we found no other audiovisual saliency model for complex dynamic scenes saliency computation. Thus, for the purpose of comparison with state-of-the-art methods, in this study, we compared our proposed audiovisual saliency model against two state-of-the-art visual saliency models. The first model proposed in [26] derives pixel-level spatial/temporal saliency map from superpixel-level spatial/temporal saliency map constructed using motion and color histogram features. The other spatio-temporal saliency detection model proposed by Liu et al. [27] is based upon superpixel-level graph and temporal propagation.

TABLE 4. Experimental results.

	Ours	Liu et al. 2014 [26]	Liu et al. 2016 [27]
AUC	**0.739**	0.716	0.712
D_{KL}	**4.153**	4.255	6.437
NSS	0.913	1.091	**1.139**
CC	0.147	**0.165**	0.161

For evaluation, We computed three saliency maps for the selected videos from DIEM dataset using the aforementioned two state-of-the-art methods and our proposed model. Using the evaluation measures described in Section IV-C, average scores for the resulting saliency maps were compared to assess the eye movements predictability of the proposed model.

V. RESULTS

The proposed model is compared against two state-of-the-art methods [26], [27] on 25 random videos using first 300 frames per video from DIEM dataset. Table 4 shows averaged evaluation scores over all the videos for the three models: the two comparison methods and the proposed solution.

It can be seen from Table 4 that the proposed model not only outperforms both comparison methods but also results in a satisfactorily higher average score in terms of AUC. Moreover, a lower D_{KL} score compared to the comparison methods indicates a better saliency model with less dissimilarity with the ground truth. For the remaining evaluation metrics, the CC and the NSS, the proposed method shows comparable performance (discussed in VI). Some of the video sequences performing better than other considering all four metrics scores are stewart_lee_1280 × 712, news_us_election_debate_1080 × 600 and one_show_1280 × 712. It is noted that, on screen sound source is a common feature in these videos with others like object occlusion, interaction etc. These results suggest that the proposed model makes better or comparable eye movement predictions, specially in scenarios where audio guides the eye movement, and thus warrant further exploration of incorporating audio features when computing spatio-temporal saliency for unconstrained videos.

Figure 4 shows saliency maps produced by the different methods. The visual comparison shows that the proposed solution performs comparatively better than the state-of-the-art methods. For instance, video sequence with an on-screen audio source-type in third row, visual models failed to correspond to the ground truth as they considered both faces salient; on the contrary, the proposed audiovisual model marked only the talking face salient.

A. TIME COMPLEXITY

Table 5 details a comparison of computation times of the proposed solution alongside the state-of-the-art methods for a 534 × 400 sized video-frame. The result shows that Liu et al. 2016 (SGSP) method [27] took the least amount of time whereas the proposed solution took the maximum time; however, excluding the audio saliency computation stage for the proposed method would result in much lower computation times than state-of-the-art methods. The table also shows a breakdown of the total time into times taken by each step of the proposed solution.

TABLE 5. Time complexity for the proposed model and state-of-the-art methods.

Method	Steps	time per frame (s)
7*Ours	Optical flow	13.406
	Segmentation	15.752
	Tracking	0.718
	Audio-Video Corr.	2.509
	Video Saliency Comp.	0.218
	Motion Map Comp.	0.078
	32.681s	
Liu et al. 2014 [26]	13.658s	
Liu et al. 2016 [27]	7.797s	

As shown in table 5 optical flow and segmentation take most of the computation time while rest of the steps are efficient enough. It is generally known that dense optical flow computation is an optimization problem and thus a compute intensive process, also the method used here estimates both forward and backward flow so that optical flow of occluded regions is also computed correctly making it more time consuming. Similarly for segmentation of diversified multiple objects meanshift is used, a non-parametric clustering technique employing kernel density estimation which is not very scalable with feature space dimension making the segmentation step compute intensive. The methods used though computationally expensive align well with the task of diversified instance segmentation and tracking, moreover, this work does not focus on the efficiency but aims at analyzing the role of audio cues in predicting the saliency of complex videos.

VI. DISCUSSION

spatio-temporal saliency detection is a challenging problem, existing research though doing well generally may fail in scenarios where attention is driven by other stimuli in addition to visual cues, and hence computation of a saliency model based upon multiple stimuli has been proposed [22], [31], [32]. Yet existing work in the field of spatio-temporal saliency modeling lacks in a generic solution incorporating both audio and video stimuli, it is either close to being non-existent or limited to some specific categories of videos.

A major reason for this lack in literature may be due to one of the foremost challenges of audiovisual saliency models; that is, the localization of audio source in a given frame. Many of the localization methods do not do well in the case of dynamic videos, as they assume there to be a single audio source. These methods either use microphone arrays

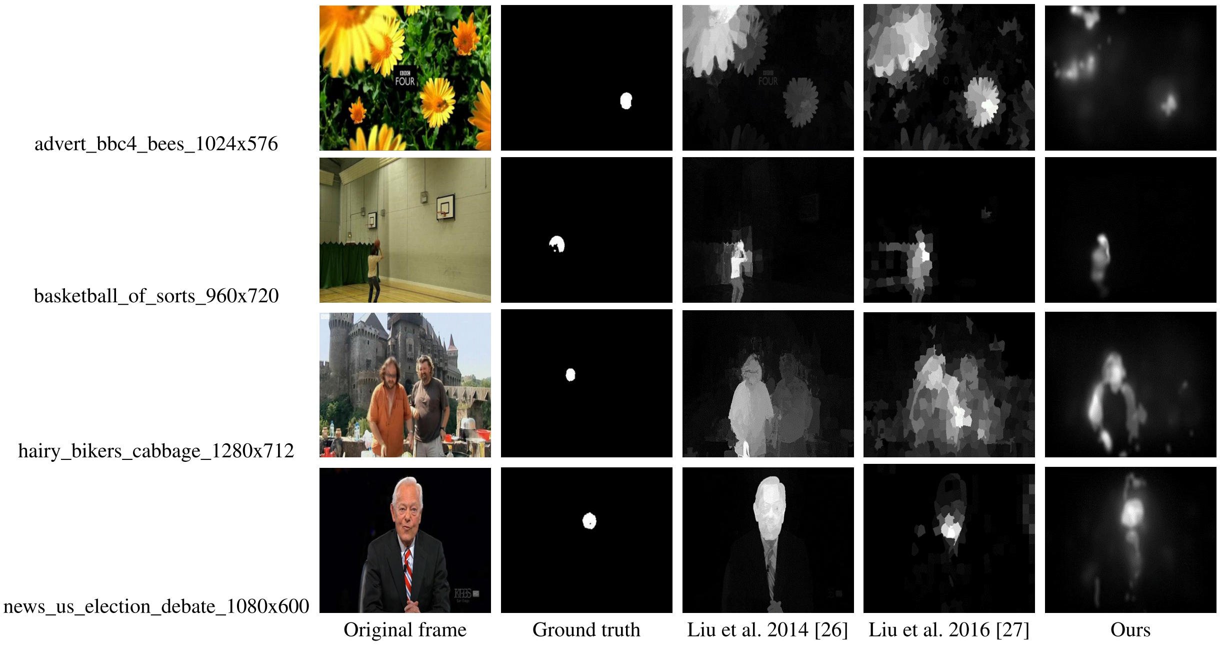
advert_bbc4_bees_1024x576
basketball_of_sorts_960x720
hairy_bikers_cabbage_1280x712
news_us_election_debate_1080x600
Original frame
Ground truth
Liu et al. 2014 [26]
Liu et al. 2016 [27]
Ours

FIGURE 4. Comparison of our results with other methods against the ground-truth on DIEM dataset.

to triangulate this single source or only target stationary sources in a scene. On the other hand, methods exempt from these restrictions using correlation analysis between audio and video segment the audio source as a set of some relevant pixels rather than an object.

In more recent works object segmentation precedes audio-visual correlation making audio source separation maintain the source object's shape. But both audio and video signals being from different domains, reliable correlation requires a transformation of the features into a suitable space. Also in addition to devising a method to relate an audio signal descriptor with object descriptors in a video frame, segmentation and tracking of all the diversified objects in a video frame is in itself a challenging task. And the literature also lacks in such techniques for multiple objects aimed for video sequences with diverse objects, which is the case in our dataset with no a priori information about object properties like shape, color, size, etc.

It is worth mentioning here that existing spatio-temporal saliency models completely ignore the audio signal present in the input media. However, a number of experimental studies [16], [17], [47] discuss the influence of aural stimuli on early attention when viewing complex scenes; that is, audio stimuli can provide useful information in guiding eye movements. This influence can be incorporated into existing bottom-up models by the inclusion of low-level audio properties like energy, frequency, amplitude, etc. The resulting audiovisual saliency model makes more sense to be used in application areas like movie summarization/compression, event detection, gaze prediction, and robotic vision and interaction.

The knowledge gap thus created should be fulfilled as the need for a general purpose solution is evident from the use of saliency detection in various fields like video summarization, compression, robotic vision, etc. The solution proposed in this work is an attempt to bridge this knowledge gap and the results presented in V via saliency metrics support that the proposed audiovisual saliency model performs better by more accurately predicting eye fixations than existing visual saliency models.

In terms of eye movement predictability, in comparison to the state-of-the-art spatio-temporal and audiovisual models, the proposed audiovisual saliency model performed better for two evaluation metrics; however, resulted in comparable scores for the other two metrics. The result can be attributed to the difficulty in segmenting and tracking of multiple interacting objects in varying conditions like motion blur, crowd, etc. Moreover, multiple and/or off-screen audio sources make it a more challenging task to locate the audio source [48], in consequence, affecting the model's performance.

The proposed saliency model exhibits higher time complexity apparently due to dense optical flow computation, which is inherently compute-intensive being an optimization problem. The main advantage of using the method is that it estimates both forward and backward flow, and hence, optical flow of occluded regions is also computed correctly. Other alternative motion estimation approaches are block-matching and phase correlation that can be used instead to propose a more efficient solution. Likewise, segmentation of multiple objects is also a complex task involving mean-shift segmentation, a non-parametric clustering using kernel density estimation that is not very scalable with feature space dimension. All these aforementioned steps make color image segmentation highly compute intensive. Alternatively, simpler histogram or superpixel-based segmentation methods can be used to reduce computational complexity, as well as, increasing the model's predictability.

A shortcoming of the current study is use of a subset of the available dataset for evaluation of the proposed audiovisual model. It may be interesting to perform the evaluation using the entire dataset and/or other available datasets to enforce our findings that aural stimuli can provide useful information in guiding eye movements alongside visual stimuli. However, there are not many datasets available designed for the purpose of audio-visual saliency prediction, we intend to collect a new dataset and evaluate the method on it in future. All in all, the proposed solution scored reasonably well; however, further improvements can be made. For example, improving segmentation and tracking techniques may contribute toward a better audio saliency map, and in turn towards a better final saliency map. Furthermore, the use of a more sophisticated visual saliency model, as well as the use of more suitable combination techniques, can improve the final result.

VII. CONCLUSION

Existing bottom-up saliency models only use visual stimuli while the audio stimuli present in the input media remain unused. In this paper, we proposed an audiovisual saliency model incorporating both low-level visual and audio information to produce three different saliency maps: an audio saliency map, a motion saliency map, and a visual saliency map. These resulting saliency maps were linearly combined with the audio saliency map weighted twice compared to the other two saliency maps to get a final saliency map. The final saliency maps produced by the model were evaluated for the DIEM dataset using four different evaluation criteria. The results show an overall improvement against two state-of-the-art visual saliency models and enforce the idea that aural stimuli can provide helpful information to guide eye movements. In future, we plan to collect a new dataset for the audio-visual saliency prediction task in uncategorized videos and also work on improving the efficiency of the proposed method by incorporating less compute intensive video instance segmentation techniques. Furthermore, this work specifically focused on visual saliency computation methods for comparison as the intention was to test the hypothesis of aural signals being impactful or not for video saliency computation, next we intend to compare the solution with state of the art audio-visual methods.

REFERENCES

[1] M. Tliba, M. A. Kerkouri, B. Ghariba, A. Chetouani, A. Coltekin, M. S. Shehata, and A. Bruno, "SATSal: A multi-level self-attention based architecture for visual saliency prediction," *IEEE Access*, vol. 10, pp. 20701–20713, 2022.

[2] L. Niu, L. Aha, J. Mattila, A. Gotchev, and E. Ruiz, "A stereoscopic eye-in-hand vision system for remote handling in ITER," *Fusion Eng. Des.*, vol. 146, pp. 1790–1795, Sep. 2019.

[3] S. Nousias, G. Arvanitis, A. S. Lalos, G. Pavlidis, C. Koulamas, A. Kalogeras, and K. Moustakas, "A saliency aware CNN-based 3D model simplification and compression framework for remote inspection of heritage sites," *IEEE Access*, vol. 8, pp. 169982–170001, 2020.

[4] Q. Yao and X. Gong, "Saliency guided self-attention network for weakly and semi-supervised semantic segmentation," *IEEE Access*, vol. 8, pp. 14413–14423, 2020.

[5] Y. Jones, F. Deligianni, and J. Dalton, "Improving ECG classification interpretability using saliency maps," in *Proc. IEEE 20th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2020, pp. 675–682.

[6] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 155–162.

[7] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[8] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. CVPR*, Jun. 2011, pp. 433–440.

[9] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 241–248.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.

[11] S. Marat, A. Rahman, D. Pellerin, N. Guyader, and D. Houzet, "Improving visual saliency by adding 'face feature map' and 'center bias,'" *Cognit. Comput.*, vol. 5, no. 1, pp. 63–75, Mar. 2013.

[12] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, 2007.

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

[14] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 231–243, May 2009.

[15] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2012.

[16] C. Quigley, S. Onat, S. Harding, M. Cooke, and P. König, "Audio-visual integration during overt visual attention," *J. Eye Movement Res.*, vol. 1, no. 2, pp. 1–17, Sep. 2008.

[17] G. Song, D. Pellerin, and L. Granjon, "Different types of sounds influence gaze differently in videos," *J. Eye Movement Res.*, vol. 6, no. 4, pp. 1–13, Oct. 2013.

[18] J. Chen, H. Song, K. Zhang, B. Liu, and Q. Liu, "Video saliency prediction using enhanced spatiotemporal alignment network," *Pattern Recognit.*, vol. 109, Jan. 2021, Art. no. 107615.

[19] F. Yan, C. Chen, P. Xiao, S. Qi, Z. Wang, and R. Xiao, "Review of visual saliency prediction: Development process from neurobiological basis to deep models," *Appl. Sci.*, vol. 12, no. 1, p. 309, Dec. 2021.

[20] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, and S. Kollias, "Spatiotemporal saliency for video classification," *Signal Process., Image Commun.*, vol. 24, no. 7, pp. 557–571, Aug. 2009.

[21] A. Coutrot and N. Guyader, "Multimodal saliency models for videos," in *From Human Attention to Computational Attention* Cham, Switzerland: Springer, 2016, pp. 291–304.

[22] N. Ould-Sidaty, M.-C. Larabi, and A. Saadane, "An audiovisual saliency model for conferencing and conversation videos," in *Proc. Image Quality Syst. Perform. (IQSP), Electron. Imag. Symp.*, 2016, pp. 1–6.

[23] A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Mach. Vis. Appl.*, vol. 22, no. 1, pp. 61–76, 2011.

[24] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 693–708, Apr. 2010.

[25] S. Marat, M. Guironnet, and D. Pellerin, "Video summarization using a visual attention model," in *Proc. 15th Eur. Signal Process. Conf.*, Sep. 2007, pp. 1784–1788.

[26] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.

[27] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, Dec. 2017.

[28] E. Birmingham, W. F. Bischof, and A. Kingstone, "Saliency does not account for fixations to eyes within social scenes," *Vis. Res.*, vol. 49, no. 24, pp. 2992–3000, Dec. 2009.

[29] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *J. Vis.*, vol. 11, no. 5, pp. 1–5, May 2011.

[30] Z. Kapoula and E. Pain, “Differential impact of sound on saccades vergence and combined eye movements: A multiple case study,” *J. Clin. Stud. Med. Case Rep.*, vol. 5, p. 95, Jan. 2020.

[31] K. Rapantzikos, G. Evangelopoulos, P. Maragos, and Y. Avrithis, “An audio-visual saliency model for movie summarization,” in *Proc. IEEE 9th Workshop Multimedia Signal Process.*, Oct. 2007, pp. 320–323.

[32] A. Coutrot and N. Guyader, “An audiovisual attention model for natural conversation scenes,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1100–1104.

[33] 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.

[34] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[35] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, “Fast and efficient saliency detection using sparse sampling and kernel density estimation,” in *Proc. Scand. Conf. Image Anal.* Cham, Switzerland: Springer, 2011, pp. 666–675.

[36] X. Min, G. Zhai, J. Zhou, X. P. Zhang, X. Yang, and X. Guan, “A multimodal saliency model for videos with high audio-visual correspondence,” *IEEE Trans. Image Process.*, vol. 29, pp. 3805–3819, 2020.

[37] K. Li, J. Ye, and K. A. Hua, “What’s making that sound?” in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 147–156.

[38] D. Liu, Y. Cui, W. Tan, and Y. Chen, “SG-Net: Spatial granularity network for one-stage video instance segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9816–9825.

[39] K. Zhang, Z. Zhao, D. Liu, Q. Liu, and B. Liu, “Deep transport network for unsupervised video object segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8781–8790.

[40] H.-S. Chang and Y.-C.-F. Wang, “Superpixel-based large displacement optical flow,” in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 3835–3839.

[41] Z. Song, Y. Wang, J. Fan, T. Tan, and Z. Zhang, “Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 3222–3231.

[42] J. Yagnik, D. Strelow, D. A. Ross, and R.-S. Lin, “The power of comparative reasoning,” in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2431–2438.

[43] D. Bradley and G. Roth, “Adaptive thresholding using the integral image,” *J. Graph. Tools*, vol. 12, no. 2, pp. 13–21, 2007.

[44] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, “Clustering of gaze during dynamic scene viewing is predicted by motion,” *Cogn. Comput.*, vol. 3, no. 1, pp. 5–24, 2011.

[45] T. Judd, F. Durand, and A. Torralba, “A benchmark of computational models of saliency to predict human fixations,” MIT, Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, 2012.

[46] R. J. Peters, A. Iyer, L. Itti, and C. Koch, “Components of bottom-up gaze allocation in natural images,” *Vis. Res.*, vol. 45, no. 18, pp. 2397–2416, Aug. 2005.

[47] E. Van der Burg, C. N. Olivers, A. W. Bronkhorst, and J. Theeuwes, “Pip and pop: Nonspatial auditory signals improve spatial visual search,” *J. Exp. Psychol., Hum. Perception Perform.*, vol. 34, no. 5, p. 1053, 2008.

[48] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, “A survey of sound source localization with deep learning methods,” *J. Acoust. Soc. Amer.*, vol. 152, no. 1, pp. 107–151, Jul. 2022.

MARYAM QAMAR received the B.S. degree (Hons.) in computer science from The University of Azad Jammu and Kashmir, Muzaffarabad, Pakistan, in 2013, and the M.S. degree in computer science from the National University of Sciences and Technology, Islamabad, Pakistan, in 2017. She is currently pursuing the Ph.D. degree in artificial intelligence with Kyung Hee University, South Korea.

Since 2019, she has been a Lecturer with the Department of Computer Science and Information Technology, The University of Azad Jammu and Kashmir. Her research interests include artificial intelligence, machine learning, image and video processing, and computer vision.

SULEMAN QAMAR received the B.S. degree (Hons.) in computer science (artificial intelligent) and the M.S. degree (Hons.) in computer science (deep reinforcement learning).

He is currently working as a Research Assistant with the CIPMA Laboratory, Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, Pakistan. His research interests include deep reinforcement learning, deep neural networks, autonomous navigation and tracking, swarm intelligence, biomedical informatics, and medical image analysis utilising machine learning techniques.

MUHAMMAD MUNEEB received the M.Sc. degree in computer science from Khalifa University, Abu Dhabi, United Arab Emirates. He is currently working as a Research Associate with the Khalifa University of Science and Technology, under the supervision of Dr. Samuel. He works on inter-discipline problems. His research interests include algorithms, automation, genetics, medical image analysis, and optimization.

SUNG-HO BAE (Member, IEEE) received the dual B.S. degrees (summa cum laude) in electrical engineering and computer science from Kyung Hee University, South Korea, in 2011, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012 and 2016, respectively, under the supervision of Prof. Munchurl Kim. From 2016 to 2017, he was a Postdoctoral Associate with the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Boston, MA, USA, under the supervision of Prof. Wojciech Matusik. Since 2017, he has been an Assistant Professor with the Department of Computer Science and Engineering (CSE), Kyung Hee University. He is also working as the Dean of the Department of CSE. His research interests include inverse problems in image processing, neural architecture search, model compression, and robustness of deep neural networks. He was a recipient of the Microsoft Research Asia Fellowship Nomination Award, in 2014, the Silver Prize of Samsung Electronics Paper Award, in 2015, the Qualcomm Innovation Awards, in 2013, 2015, and 2016, and the Excellent Research Achievement Awards from KAIST, in 2014 and 2015.

ANIS RAHMAN received the master’s degree in parallel and distributed systems from Joseph Fourier University, France, and the Ph.D. degree in computer science from Grenoble Alpes University, France, in 2013. He is currently an Assistant Professor with the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Pakistan. His main research interests include modeling of visual attention by assessing the different mechanisms guiding it, salient multi-object image and video segmentation and tracking, and efficient implementations of large-scale scientific problems on commodity graphical processing units.