DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-217

An Audiovisual Saliency Model For Conferencing and Conversation Videos

Naty Ould Sidaty[1] ***, Mohamed-Chaker Larabi***[1] ***, Abdelhakim Saadane***[2]
[1] ***XLIM Lab., University of Poitiers, France***
[2] ***XLIM Lab., Polytech, University of Nantes, France***

Abstract

Visual attention modeling is a very active research area. During the last decade several image and video attention models have been proposed. Unfortunately, the majority of classical video attention models do not take into account the multimodal aspect of the video (visual and auditory cues). However, several studies have proven that human gazes are affected by the presence of the soundtrack. In this paper we propose an audiovisual saliency model that can predict the human gaze maps when exploring a conferencing or conversation videos. The model is based on the fusion of spatial, temporal and auditory attentional maps. Thanks to a real-time audiovisual speaker localization method, the proposed auditory maps are modulated by the enhanced saliency region of speakers compared to the other faces in the video. Classical visual attention measures have been used to compare the predicted saliency maps with the eye-tracking ground truth. Results of the proposed approach, using several fusion methods, show a good performance whatever the used spatial models.

Introduction

Visual attention is a selective process and a clever mechanism of the human visual system (HVS). It allows selecting the most attractive areas of the visual scene, called saliency regions. The visual attention studies permit to use this HVS property in various applications: computer vision (recognition and objects detection, tracking, compression ...), computer graphics (image rendering, dynamic lighting, ...) and robotics. On the research side, efforts have been devoted on studying and modeling of visual attention and numerous image and video saliency models have been proposed ([1], for more details). The majority of existing saliency models do not take into account the multimodal aspect of the video (audio and image). Consequently, audio has never been considered as a real attractive feature. However, experimental studies in the multimedia field have demonstrated that human perception is influenced by both modalities audio and video [12]. Furthermore, in TVs and videoconferencing applications, faces and particularly talking faces are known for their attractive character [2]. So far, an audio visual saliency model that takes into account the attributes mentioned above is of paramount importance.

The objective of this work is to propose an audiovisual saliency model that can predict the human gaze maps when exploring conferencing or conversation videos. The main idea of the proposed approach is to detect and localize the speakers face and strengthen its visual saliency compared to the other faces present in the video.

Eye-tracking Experiment

As mentioned in the introduction, visual and auditory information have a major influence on human perception. Auditory signals can influence our visual perception and vice versa. In this section we take advantage of analysing the influence of audio signal on the eye movements of observers when showing video sequences. Regarding still images, this influence has been widely discussed in the literature. The authors of [3] have realised a set of psycho-physical tests to measure the influence of sound on the human visual attention. Thus, a still image with a sound source located at the top, bottom, left and right on this image was presented to the participants. Eye movements associated with this experiment were recorded in three test conditions: auditory, visual and audiovisual. They found that eye movements of participants in audiovisual conditions are spatially biased towards the region of the image corresponding to the sound source. However, the influence of the sound information for dynamic content has never been addressed. The almost all eye tracking experiments are performed on original video without sound component. Observers are often invited to watch silent movies, which is far from the reality of the use of video content where auditory signals are usually come coupled with visual signals. In this paper, we are interested particularly for video conference applications, where audio signals are of paramount importance.

In order to investigate the influence of audio signals on visual attention, we perform an Eye-tracking experiment in which eye movements of participants, looking at a set of video sequences, are recorded. To quantify this influence, videos are presented in two conditions: visual (without their soundtrack) and audiovisual (with their soundtrack). We initially describe the used video database, where videos imitating video conference scenes were recorded. Then, we study the importance of faces and especially talking heads in the video

studied the importance of faces and especially talking heads in the video. Then, we describe the used video database, where videos imitating video conference scenes were recorded. Finally, we describe the conditions for achieving our eye tracking experiments and we present the results obtained by analysing the ocular dispersion of participants in visual and audiovisual conditions.

DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-217

Stimuli

Lacking access to an existing database, that can mimic the video conferencing applications, we have been induced to create our database in XLIM Lab., named XLIMedia. The videos of these applications consist mainly of a number of people interacting and taking spontaneous speech. Five conversation scenarios are considered:

- The sound source is coming from the outside (narrator).
- One person talks among several: a single speech signal.
- Two people speak alternately: two speech signals.
- Two people speak simultaneously: two speech signals.
- Many people (>2) talking at once: different speech signals.

Videos are recorded in such a way that the region that produces the sound (the talking head) changes from video to another. A camcorder Sony HVR-VE1 equipped with a stereo microphone is used for this purpose. Finally, five categories of videos, named OutsideTalk, SinglTalk, AlterTalk, SimulTalk and AllTalk, have been used in the experiment. Videos in the final category have been taken from CUAVE database [20, 21]

Apparatus

For the conducted experiments, we used the binocular eye-tracker Tobii TX-120 allowing to track both eyes simultaneously. Before the eye-tracking experiment, the 9-point calibration procedure is performed for each user for an accurate recording.

The accuracy of Tobii TX-120 is equal to 0.5° and the data is delivered every 8 ms.

Participants

Fifteen healthy naive subjects have been invited from the University of Poitiers to participate to the eye-tracking experiments for free. There were 11 male and 4 female with ages ranging between 21 and 43. Visual acuity has been checked by FrACT (Freiburg Visual Acuity Test) and it was around 1 with or without correction. All subjects have been screened using the Ishihara compatible color vision test for detecting color blindness.

Procedure

The procedure in this eye-tracking experiment is shown in Figure 1. First, the stimuli were presented without audio. A gray image with a message "Now listen carefully" was displayed before presenting the same stimuli accompanied by their original soundtrack. Each stimulus, with or without sound, has a total duration of 57 seconds. The gray image has a duration of approximately five seconds. Two videos with and without sound, were used as test sequences to familiarize participants with the process of eye tracking tests. Recorded data of these sequences are not included in the results analysis phase.

At the end of this experience, each video is viewed by seventeen participants in both conditions: visual and audiovisual. All data related to eye movement (fixations, saccades, etc.) are taken chronologically in a single file (csv) with the TobiiStudio software.

Figure 1. *Experiment procedure of visualisation*

Data processing

The conducted eye-tracking experiments provided information about the eye-movements of both eyes. Three types of data have been recorded: fixations, saccades and blinks. In our study, only information about observer's eye fixations are used. After obtaining these fixations for each frame of the video, the fixations of the average subject have been calculated. For this, fixations from different records are accumulated and each point is normalized by the number of participants. So far, the value of each pixel of the map demonstrates its ability to attract attention. It means that the higher this value is, the more attractive this area will be.

The fixation map FM^k for each observer k is then constructed (1) and the final fixation map FMM is the mean of these singular fixation maps, as given in the equation 2, where N is the number of observers.

$$FM^k(x,y) = \sum_{i=1}^{M} \delta(x-x_i, y-y_i) \tag{1}$$

$$FMM(x,y) = \frac{1}{N} \sum_{k=1}^{N} FM^k(x,y) \tag{2}$$

However, it seems obvious that the eye does not fix at a particular point on an image but on an area with visual size close to that of the fovea. Hence, the fixation density map, for each frame, is obtained by convolution of FM with a bi-dimensional Gaussian filter g_{σ_x,σ_y}, with a standard deviation equal to 1, as given by the equation 3. The choice of this value is based on the assumption that the fovea can covers between 1 and 2 degrees of visual angle.

$$FMM(x,y) = FMM(x,y) * g_{\sigma_x,\sigma_y}(x,y) \tag{3}$$

where g_{σ_x,σ_y} is a bi-dimensional Gaussian filter given by equation 4 below:

$$g_{\sigma_x,\sigma_y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} * \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right) \tag{4}$$

We end up with two ground truth: without audio, constructed using human fixations in visual condition and without audio, constructed using human fixations in audiovisual condition (stimuli presented with their soundtrack). Figure represents an example frame of obtained ground truth with and without audio for the same original frame (sequence SinglTalk).

Evaluation measures

The objective of a visual saliency model is to predict the most attractive regions of a scene as do human visual system. To check

DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-217

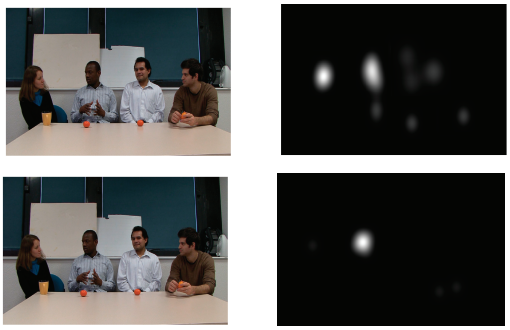

Figure 2. *Example frame of the two constructed ground truth, top without audio and bottom with audio*

the performance of a model, data from the ground truth as well as validation metrics are required. The performance of a saliency model is measured by comparing its predicted areas with the regions really fixed by observers during the eye-tracker experiment. Different similarity measures are widely used in the literature to assess this performance: Correlation coefficient (CC) [16], Area Under the ROC curve (AUC) [17, 18], Kullback-Leibler Divergence (KLD) [16, 18], Normalized scanpath Saliency (NSS) [19], Earth Mover's Distance (EMD), String Editing Distance (SED), etc.

As part of our work, four of these measures known to be effective [5], namely the CC, AUC, KLD and NSS are used. These four methods are used, individually or jointly, to assess performance in most saliency models of literature .

Proposed audiovisual saliency model

The proposed audiovisual saliency model is based on the fusion of three saliency maps: spatial, temporal and auditory. First, we have performed a comparison between the state-of-the-art 2D static saliency models on our eye-tracking ground truths. Three well known models have been selected by their performance. Each of these models is used as the spatial component of the proposed audiovisual model. Motion, and particularly motion of objects relative to the background, is known to attract humans gaze. Therefore motion vectors have been used as temporal component of the proposed approach. In conversation and conferencing scenes the auditory cues are usually come from speaker. Consequently, the auditory map in the proposed model is based on the detected and localized talking faces. A real-time audiovisual speaker localization method has been used based on the audio-visual synchrony. Faces of the video sequences are detected and the audio energy and visual motion displacement of those faces are combined to measure the degree of synchrony. Faces having a high synchrony score are then selected to be the speaker heads. Audio saliency map is then obtained by a weighting talking heads with respect to the others. Finally a set of various fusion methods have been used to predict audiovisual salience map, as shown in figure 3. This procedure is given in detail in the following sections.

Static map

First of all we have tested the performances of a set of static saliency models from different categories that widely are used in the literature (Cognitive models, Theoretical models, Graphical

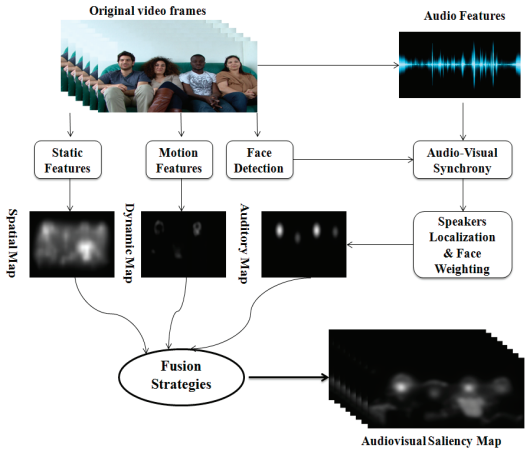

Figure 3. *Proposed audio-visual saliency model*

models, Spectral Analysis Models, etc.) [4]. Results on a well known database (MIT300) [5] are given in the following table.

Performances of all selected models on MIT300 database.

Model	Itti	Achanta	Harel	Bruce	Zhang	Nauge	Vikram	Tavakoli
CC	0.37	0.04	0.48	0.31	0.55	0.59	0.38	0.48
AUC	0.63	0.52	0.63	0.66	0.65	0.61	0.55	0.59
NSS	0.97	0.13	1.24	0.79	1.41	1.17	0.95	1.47

The obtained results show that the performance of each model greatly depend on used similarity metric. For example, Nauge model can better predict the human gaze when using CC measure, Bruce model when using AUC and Tavakoli model when using NSS measure. For that reason we decided to use several metrics for a meaningful analysis of results.

Afterwards, we have tested the performances of these models on our database. Results are shown in table 4 for the ground truth constructed in the audiovisual conditions (stimuli with soundtrack). We can clearly observe that the performances of these models have dramatically dropped when taking into account the auditory information. Despite that, we can notice that Itti's model, Harel's model and Tavakoli's model are the most efficient models on our database. In the rest of this work, we only keep these models to characterize the static view of the proposed model.

Dynamic map

Eye-tracking experiments have shown that the movement and in particular the contrast of movement is known for its attractive nature of visual attention. Therefore, the dynamic view of our proposed model is based on the characteristics of this feature, particularly on the movement of objects relative to the background.

Different approaches can be used to estimate this movement (block-matching methods, differential methods, frequency methods, etc.). The motion estimation is obtained by the resolution of an optical flow equation system. This estimate is described in details in [9, 10].

DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-217

By using the Lucas-Kanade method [11], when flow is constant in a neighborhood of q ($q = 3 \times 3 = 9$ pixels), a motion vector is then constricted. The phase of this vector corresponds to the direction of movement and the module, given by equation 5, correspond to the motion speed against the background. This module is used for the temporal saliency map for each frame of the video. We used the algorithm proposed by [9] to generate these temporal saliency maps.

$$A = \sqrt{V_x^2 + V_y^2} \tag{5}$$

After having build temporal maps we have performed a simple fusion of static and temporal view. All maps are normalised and the final spatio-temporal maps C_{Final} are given by equation 6. Table 2 illustrates the performances of spatio-temporal models on XLIMedia database using three efficient models as static view. In blue the performances are better without integrating the dynamic view. These results show that spatio-temporal models are not applicable alone in our context of study. This can be explained by the very low degree of motion in videoconferencing applications. An example frame with its corresponding dynamic map is given in figure 4 where bright areas correspond to the motion regions in the video.

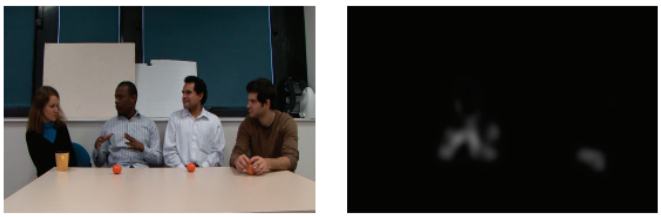

Figure 4. *Example frame with its temporal map*

$$C_{Final} = \frac{1}{2}(S_m + T_m) \tag{6}$$

Performances of spatio-temporal models on our XLIMedia database (ground truth with audio)

Static model	Harel et al.				Itti et al.				Tavakoli et al			
Measure	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS
OutsideTalk	0.20	0.18	2.17	0.84	0.20	0.20	2.29	0.90	**0.29**	**0.21**	**1.94**	**0.99**
SinglTalk	0.23	0.23	2.54	1.12	0.19	0.14	2.69	1.00	0.24	**0.24**	2.35	1.22
AlterTalk	0.20	0.19	2.72	1.07	0.21	0.21	2.78	1.17	0.24	**0.26**	2.42	1.22
SimulTalk	0.17	**0.19**	2.48	0.93	0.20	0.18	2.47	1.09	**0.29**	0.15	**1.97**	**1.36**
AllTalk	0.21	**0.27**	2.11	0.82	0.26	0.24	2.06	**0.93**	0.17	0.22	2.17	0.66

Auditory map

As we have mentioned above, the semantic content, although it has been proved in the 'visual' case, is also of major importance for observers when exploring of an audiovisual scene. Objects, faces and more specifically the speakers are known for their attraction relative to the rest of the content [12]. The talking heads attract more attention in different applications (video conferencing, meetings, TV shows, etc.). They were treated as " silent faces" in most models of literature, integrating face component. The final saliency map is obtained by assigning equal importance to all the faces in the video. To cope with this inconvenience we have proposed to integrate an auditory view based on the location of talking heads in the video sequence.

First of all, faces and soundtrack of the video are extracted and processed separately. In order to locate an active speaker (for each time window) an audiovisual synchronisation method have been used [13]. Here the audio signal, a(t), is represented by the acoustic energy contained in each frame of the video. The visual signals are represented by the "faces" stream presents in the video. In fact, each face is extracted and only the lower halves, contained the mouth region, are used. Each of them are subdivided into macroblocks (MB). For each MB, the motion feature $f_{vn}(t)$ is computed based on the Block Matching Algorithm [14]. The representation of these audio-visual features are illustrated in figure 5.

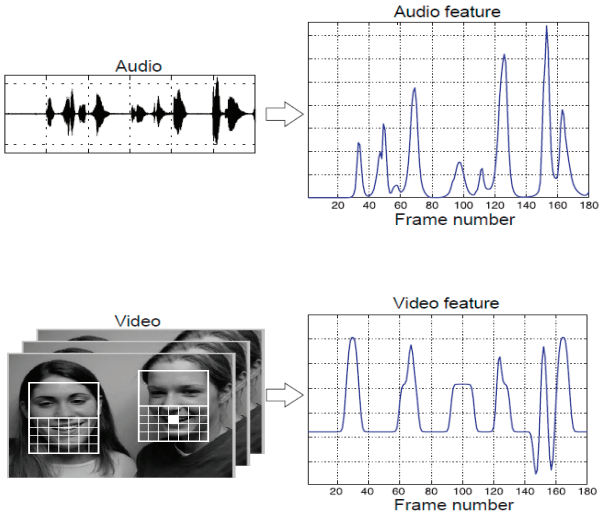

Figure 5. *Audio and visuals signal representation. Image extracted from [13]*

We end up with one representation vector for the audio and N representation vectors for each MB of the processed face. The pics of each representation indicate the presence of a phenomenon (sound for audio vector and lips movement of mouth region). An activation vectors based on these peaks locations are then constructed. These vectors equal 1 where peaks occur and 0 otherwise. The scalar product S_n between these activation vectors gives an estimate of the degree of synchrony between the audio and the motion features as given by equation 7. The active speaker (for each time window) has the highest synchronization score S_k.

$$S_k = \sum_{n=0}^{N-1} S_n, \quad where: S_n = \langle y_a(t), y_{v_n}(t) \rangle \tag{7}$$

We have tested this speaker localisation method on our video database. Results therefore indicate a high percentage of good detection, as illustrate in the following table.

After having located speaker's and non speaker's region, the auditory maps A_m are then construed. They are obtained by applying Gaussian kernels for each detected face. Talking faces are

DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-217

Sequence	Number of frame	False Detection [image]	Accuracy
SiglTalk	224	26	88.39 %
AlterTalk	305	22	92.78 %
SimulTalk	360	16	95.67 %
AllTalk	217	14	93.54 %
Mean			**92.59** %

weighted according to the Eye-tracking experiment (0.7 for talking faces and 0.3 for salient faces). An example of an auditory map is represented by figure 6.

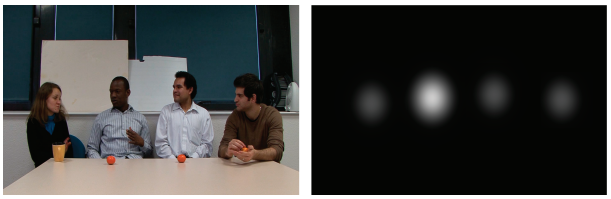

Figure 6. *An exemple frame with its corresponding auditory map when one face talk*

We have performed a simple fusion method as it was done above. static, dynamic and auditory maps are normalised and the fused into a single card as given by equation 8. Results of this fusion are given in table 3 inadequate that performances of spatio-temporal models are clearly improved by adding our proposed auditory view. These results are graphically represented in figure 7 for different similarity metrics.

$$C_{Final} = \frac{1}{3}(S_m + T_m + A_m) \tag{8}$$

Performance of audiovisual saliency model on XLIMedia database.

Performance: $CC \to \pm 1, AUC \to 1, KLD \to 0, NSS \to +\infty$

static model	Harel at al.				Itti et al.				Tavakoli et al.			
Measure	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS
Mean	0.58	0.64	1.09	2.32	0.60	0.63	1.13	2.39	0.64	0.64	0.85	2.49
S. deviation	0.09	0.09	0.28	0.26	0.08	0.13	0.29	0.24	0.14	0.11	0.33	0.42

After showing the importance of auditory cues to increasing the model performance, we have tested some fusion methods that are widely used in the literature, these methods are briefly described below:

1-**Normalization and Sum NS**: Normalization of all the maps (SM) to the same dynamic range (between 0 and 1), the final saliency map C_{Final} is given by equation 9.

$$C_{Final} = \frac{1}{\sum j}(\sum_i \mathcal{N}(SM_i)) \tag{9}$$

where $\mathcal{N}$ is the normalization operator and j the number of view (here j=3).

2-**Normalization and Maximization NM:** Maximum operator instead of summation (compare to NS)

$$C_{Final} = \max_j \mathcal{N}(SM_i) \tag{10}$$

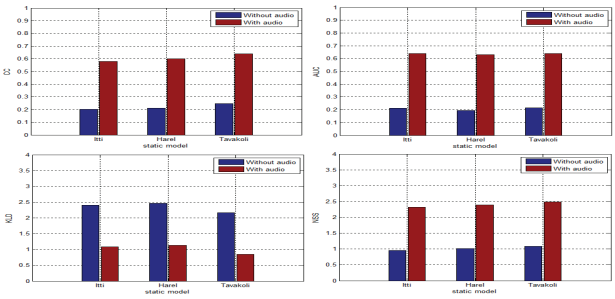

Figure 7. *Performances of proposed audiovisual saliency model after having integrate auditory component*

The sole advantage of these first two methods refers to their simplicity.

3- **Coherent Normalization and Sum CNS:** Normalization of saliency maps by an empirical value for each visual dimension. Particular pictures with high contrast color are used for the choice of this empirical value. This method is used in order to take into account the relative importance between maps.

$$C_{Final} = \frac{1}{\sum j}(\sum_i \mathcal{N}_c(SM_i)) \tag{11}$$

4-**Coherent Normalization Sum and Product CNSP:**

$$C_{Final} = \sum_j \mathcal{N}_c(SM_i) + \prod_j (1 + \mathcal{N}_c(SM_i) \tag{12}$$

These last two methods are used in order to promoting the maps having few saliency peaks and in removing the maps having an uniform distribution and a lot of saliency peaks [15].

5- **Global Non-Linear Normalization followed by Normalization GNLNS:**

$$C_{Final} = \sum_j [(\mathcal{N}(SM_i)(M_i - m_i)^2] \tag{13}$$

Results of these five fusion methods are represented in table 5. We can notice that the performance of models not only depends on the used metric but also on the fusion method. These results show that the GNLNS give the best performance compared to the other employed methods.

Conclusion

Nearly all existing visual attention models do not take into account the auditory cues. Few of the models that use face information do not take a distinction between the talking heads and dumb. Experiment results have shown that viewers were attracted by the speakers face more than the others. In this work we proposed a real-time audiovisual saliency model that highlighting this salient information. According to results, the integration of an audio component allows to significantly improve the performance of these spatio-temporal saliency models in the context of video-conferencing applications. In addition, the use of static models as a spatial view of our approach does not change significantly, except for KLD measure, the global performance of the proposed model.

DOI: 10.2352/ISSN.2470-1173.2016.13.IQSP-217

Performance of selected models on our XLIMedia database (ground truth with audio)

Model	Achanta et al.				Harel et al.				Itti et al.				Tavakoli et al.			
Measure	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS
Mean	0.04	0.04	3.75	0.25	0.20	**0.28**	2.46	1.02	0.19	0.20	2.53	**1.15**	**0.23**	0.24	**2.26**	1.11

Model	Bruce et al.				Zhang et al.				Vikram et al.				Nauge et al.			
Measure	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS
Mean	0.14	0.20	3.18	0.69	0.15	0.17	3.17	0.67	0.11	0.12	2.83	0.78	0.09	0.12	3.06	0.76

Performance of the proposed model using several fusion methods

Fusion method	NS				NM				CNS				CNSP				GNLNS			
Measure	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS	CC	AUC	KLD	NSS
Itti model	0.60	0.63	1.13	2.39	0.52	0.55	1.35	2.16	0.61	0.65	0.93	2.63	0.66	0.66	0.87	2.85	0.74	0.74	0.67	3.11
Harel model	0.58	0.64	1.09	2.32	0.52	0.54	1.28	2.13	0.61	0.60	0.88	2.62	0.66	0.70	0.82	2.83	0.74	0.74	0.62	3.10
Tavakoli model	0.64	0.64	0.85	2.49	0.56	0.53	1.01	2.28	0.65	0.67	0.60	2.76	0.71	0.75	0.39	2.98	0.75	0.76	0.48	3.17

References

[1] A. Borji and L. Itti. State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):185207, (2013).

[2] M. Bendris, D. Charlet and G. Chollet, Lip Activity Detection For Talking Faces Classification In Tv-Content, International Conference on Machine Vision (2010).

[3] C. Quigley, S. Onat, S. Harding, M. Cooke and P. Konig, Audio-visual integration during overt visual attention, Journal of Eye Movement Research, (2008).

[4] A. Borji, D.N. Sihite and Laurent Itti, Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 1, (2013).

[5] T. Judd, F. Durand and A. Torralba, A Benchmark of Computational Models of Saliency to Predict Human Fixations, MIT Technical Report, (2012).

[6] L. Itti, C. Koch, and E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, Nov. (1998).

[7] J. Harel, C. Koch, and P. Perona, Graph-Based Visual Saliency, Neural Information Processing Systems, vol. 19, pp. 545-552, (2006).

[8] H.R. Tavakoli, E. Rahtu, and J. Heikkilä, Fast and efficient saliency detection using sparse sampling and kernel density estimation, in Proc. 17th Scandin. Conf. Image Anal., pp. 666675, (2014).

[9] Sophie Marat, Modles de saillance visuelle par fusion dinformations sur la luminance, le mouvement et les visages pour la prdiction de mouvement oculaire lors de lexploration de vidos. Thse de doctorat, Institut polytechnique de Grenoble (2010).

[10] E. Bruno, De l'estimation locale l'estimation globale de mouvement dans les squences d'images. Thse de doctorat, Institut polytechnique de Grenoble, (2001).

[11] A. Bruhn, J. Weickert and C. Schnrr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. International Journal of Computer Vision, 61(3):211231, (2005).

[12] N. Sidaty, M-C. Larabi et A. Saadane, Towards Understanding and Modeling Audiovisual Saliency Based on Talking Faces, 10th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Marrakech (2014).

[13] G. Monaci, Towards Real-Time audiovisual speaker localization. 19th European Signal Processing Conference, EUSIPCO, Barcelona (2011).

[14] Y. Nie and K.K. Ma. Adaptive rood pattern search for fast block-matching motion estimation. IEEE Trans. Image Processing, 11(12):14421449, (2002).

[15] C. Chamaret, J-C. Chevet, and O. Le Meur, Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies, In Proceedings of 17th IEEE International Conference on Image Processing (ICIP), (2010).

[16] B. W Tatler, R. J. Baddeley and I. D Gilchrist, Visual correlates of fixation selection: Effects of scale and time, Vision research, 45(5):643659, (2005).

[17] A. Borji, D. Sihite and L. Itti, Computational Modeling of Top-down Visual Attention in Interactive Environments, In British Machine Vision Conference (BMVC) (2011).

[18] U. Rajashekar, L. K Cormack and A. C Bovik, Point-of-gaze analysis reveals visual search strategies, In IS&T/SPIE Electronic Imaging Proceedings, pages 296306. International Society for Optics and Photonics, (2004).

[19] R. J. Peters, A. Iyer, L. Itti and C. Koch, Components of bottom-up gaze allocation in natural images, Vision research, 45(18):23972416, (2005).

[20] P. Besson, G. Monaci, P. Vandergheynst, and M. Kunt. Experimental framework for speaker detection on the CUAVE database. Technical Report EPFL-ITS 003, Lausanne, (2006).

[21] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy. Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus. EURASIP JASP, (11):1189 1201, (2002).