Signal Processing: Image Communication 76 (2019) 186–200

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

A behaviorally inspired fusion approach for computational audiovisual saliency modeling☆

Antigoni Tsiami [a,*], Petros Koutras [a], Athanasios Katsamanis [a], Argiro Vatakis [b], Petros Maragos [a]

[a] School of ECE, National Technical University of Athens, GR 15773, Greece
[b] Cognitive Systems Research Institute (CSRI), Greece

ARTICLE INFO

Keywords:
Audiovisual saliency
Attention
Fusion
Eye-tracking

ABSTRACT

Human attention is highly influenced by multi-modal combinations of perceived sensory information and especially audiovisual information. Although systematic behavioral experiments have provided evidence that human attention is multi-modal, most bottom-up computational attention models, namely saliency models for fixation prediction, focus on visual information, largely ignoring auditory input. In this work, we aim to bridge the gap between findings from neuroscience concerning audiovisual attention and the computational attention modeling, by creating a 2-D bottom-up audiovisual saliency model. We experiment with various fusion schemes for integrating state-of-the-art auditory and visual saliency models in a single audiovisual attention/saliency model based on behavioral findings, that we validate in two experimental levels: (1) using results from behavioral experiments aiming to reproduce the results in a mostly qualitative manner and to ensure that our modeling is in line with behavioral findings, and (2) using 6 different databases with audiovisual human eye-tracking data. For this last purpose, we have also collected eye-tracking data for two databases: ETMD, a movie database that contains highly edited videos (movie clips), and SumMe, a database that contains unstructured and unedited user videos. Experimental results indicate that our proposed audiovisual fusion schemes in most cases improve performance compared to visual-only models, without any prior knowledge of the video/audio content. Also, they can be generalized and applied to any auditory saliency model and any visual spatio-temporal saliency model.

1. Introduction

Attention can be defined as the behavioral and cognitive process of selectively concentrating on a specific aspect of information, while ignoring other perceivable input. The role of attention is vital to humans, and its mechanism has been in the research focus for many decades. A computational modeling of human attention could not only be exploited in applications like robot navigation, human–robot interaction, advertising, summarization, etc., but could also offer an additional insight in our understanding of human attention functions.

Although visual and auditory stimuli often attract attention in isolation, most of the times stimuli are multi-sensory and multi-modal, resulting in human multi-modal attention, e.g., audiovisual attention. The influence that multi-modal stimuli exert on human attention and behavior [1–4] can be perceived both in everyday life, but also through targeted behavioral experiments. It can therefore be observed that when multi-modal stimuli are incongruent they can lead to illusionary perception of the multi-modal event, as in the ventriloquist or the McGurk effect [5], while in the opposite case, where the stimuli are synchronized/aligned, they can effectively enhance both perception and performance.

In this work, we focus on investigating how multi-sensory, and specifically audiovisual stimuli can influence human bottom-up attention, namely saliency [6]. For example, in [7], a series of behavioral experiments is described, that highlights the influence of multi-modal stimuli on saliency, through an effect called "pip and pop": in a visual search task that consists of a cluttered image containing a target (that has to be identified by humans) and distractors that change dynamically, the insertion of a non-localized auditory pip synchronized with target changes can significantly enhance reaction times. It has been observed that these task-irrelevant pips make the target become more salient (i.e. "pop out"). This is just a single example of strong audiovisual interaction, the mechanisms of which have been in the focus of cognitive research for years.

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.image.2019.05.001.

* Corresponding author.
E-mail addresses: antsiami@cs.ntua.gr (A. Tsiami), pkoutras@cs.ntua.gr (P. Koutras), nkatsam@cs.ntua.gr (A. Katsamanis), maragos@cs.ntua.gr (P. Maragos).

https://doi.org/10.1016/j.image.2019.05.001
Received 28 December 2018; Received in revised form 1 April 2019; Accepted 2 May 2019
Available online xxxx

In parallel, visual and auditory saliency mechanisms have been separately well-studied, and the related findings have been integrated into individual computational models, that have already been employed in real-world applications. Some of them have been inspired and validated by behavioral experiments, like the seminal works of [8,9] for visual saliency, and [10] for auditory saliency. Motivated and validated by behavioral observations in psycho-sensory experiments these models have inspired variations and improvements, which have been used in applications like object recognition in images and prominence detection in speech.

Despite the simultaneous development of visual and auditory saliency models, few efforts have focused on creating a joint audiovisual model [11,12]. The majority of models trying to predict human attention in videos are based only on visual information excluding auditory input. On the other hand, audiovisual fusion has been found to boost performance in applications where audio and visual modalities are correlated and refer to the same event, e.g., speech recognition [13], movie summarization [14], human–robot interaction [15–17].

We aim to bridge the gap between behavioral research, where audiovisual integration has been well-investigated, and computational modeling of attention, which is mostly based on the visual modality. Our goal is to investigate ways to fuse existing visual and auditory saliency models in order to create a 2-D audiovisual saliency model that will be in line both with behavioral findings, but also with human eye-tracking data. The model should capture well the audiovisual correspondences, but its performance should not be degraded if there is no audio or if audio is not related to video. In our preliminary work [18] we introduced such an audiovisual model, based on Itti et al. and Kayser et al. and carried out some preliminary experiments to validate it through behavioral findings from a particular experiment. In this current paper we have investigated more fusion schemes in order to integrate auditory and visual saliency, we have carried out more experiments with behavioral findings but we also present an evaluation strategy that involves eye-tracking data and comparisons with various models. Some of these data have been collected for the purposes of this paper and will be publicly released. The contributions of this paper can be summarized as follows:

- Audiovisual bottom-up human attention modeling via computational audiovisual saliency modeling, inspired and validated by behavioral experiments.
- Investigation of three different audiovisual fusion schemes between visual saliency and non-localized auditory saliency, resulting in a 2-D audiovisual saliency map instead of fusion at decision or feature level. The proposed audiovisual fusion schemes for attention/saliency modeling are generic since they can be applied to any visual spatio-temporal saliency method.
- Audiovisual eye-tracking data collection for two databases, SumMe and ETMD that contain unedited user videos and highly edited movies, respectively, and unconstrained audio. The collected eye-tracking data will be released in public.
- Two-level evaluation of the proposed model:
    1. Comparison against human experimental findings from behavioral experiments in a qualitative way, aspiring to build a computational model able to explain and reproduce aspects of human attention.
    2. Comparison against human eye-tracking data from databases with audiovisual eye-tracking data and variable complexity, DIEM, AVAD, Coutrot1, Coutrot2, SumMe, and ETMD.

The rest of the paper is organized as follows: Section 2 is dedicated to an extensive review of audiovisual saliency models, behavioral findings related to audiovisual interactions, and state-of-the-art visual, and auditory saliency models . Section 3 incorporates the main aspects of computational audiovisual modeling and the proposed fusion schemes. Section 4 contains a description of the evaluation metrics as well as a detailed description of the stimuli and the conducted experiments, both for the behavioral findings and the human eye-tracking databases. Also, the newly collected audiovisual eye-tracking databases are described, and in the end of the section an analysis and discussion of the results and performance across methods and datasets is performed. The last section concludes our work.

2. Related work

Several attempts to model audiovisual attention exist in the literature, but most of them are application-specific or use spatial audio in order to fuse it with visual information.

Audiovisual attention models: Probably the first attempt in modeling audiovisual saliency appears in [19], where the eye fixation predictions in an audiovisual scene served as cues for guiding a humanoid robot. Here, the model of Itti et al. [9] is employed for visual saliency, while auditory saliency is only spatially modeled, by means of acoustic source localization. The output saliency map, via a max operation, is guided by vision unless an audio source appears in the scene.

For a similar application, in [20], the authors employ a phase-based approach for visual saliency and Bayesian surprise along with source localization for the auditory one. Salient auditory and visual events are clustered cross-modally and the audiovisual clusters' saliency is estimated by linearly combining the unimodal saliency values. In [21], the auditory saliency map is essentially the source location estimation and fusion with visual saliency is performed via a product operation. In [22], a 3D audiovisual attention model that combines visual, depth, and multi-channel audio data through two independent static and dynamic analysis paths is proposed, while in [23], an audiovisual model for videoconferencing applications is presented. It is based on the fusion of spatial, temporal, and auditory attentional maps with the latter based on real-time audiovisual speaker localization. In [24] auditory saliency is computed through acoustic event detection and visual saliency is only spatial, since the authors deal with images and not with videos.

Another model presented in [14,25] and further improved in [26], employs audiovisual saliency from a different viewpoint. The goal of this model is to predict when, and not where, audiovisual attention is drawn. All the above described models have been developed for specific applications, and their majority assumes spatial audio for auditory saliency. Also, their plausibility and validity has not been investigated through comparisons with human/behavioral data.

On the other hand, Coutrot and Guyader [12,27,28] and Song [29] have tried to more directly validate their models with humans. Among their findings is the observation that in movies, eye gaze is attracted by talking faces and music players. To match that, after estimating the visual saliency map they explicitly weigh the face image regions appropriately to generate an audiovisual saliency map to better account for eye fixations during movie viewing.

Also, Min et al. [30] in a preliminary work demonstrated that the impact of audio was up to its consistency with visual signals, while in some later works [11,31] they developed an audiovisual attention model for predicting eye fixations in scenes containing moving, sound-generating objects. For auditory attention, they employed an auditory source localization method to localize the sound-generating object on the image and fuse it with visual saliency models.

Aspiring to create a model that will not be application-specific, but built upon behavioral findings of human attention and thus generic, we essentially investigate behaviorally-inspired ways to combine existing visual and auditory saliency models. To motivate better this choice, the upper part of Fig. 1 depicts two successive frames of the previously mentioned "pip and pop" stimuli: While all lines are diagonal, there is only one, the target, that is either horizontal or vertical. All lines constantly flicker between red and green, but when the target flickers, it does so alone. Behavioral experiments have shown that when the

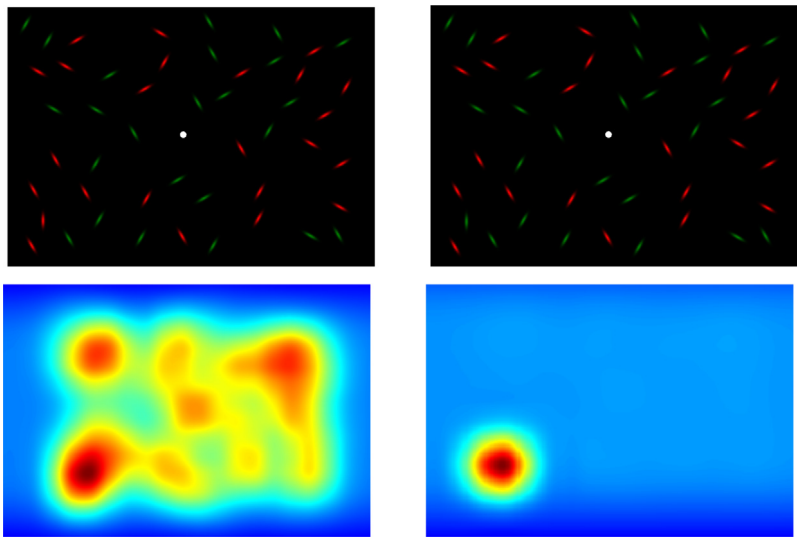

Fig. 1. The two upper figures from [7] depict the "pip & pop" stimuli during a target flicker (the vertical line in the lower left corner that flickers from red to green). Below are the visual (left) and audiovisual saliency map (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

target flicker is accompanied by a brief non-localized tone (irrelevant to the task), humans identify the target immediately compared to the visual-only case. This serves as the starting point for our model. We extensively review behavioral experiments related to audiovisual interactions, so as to extract parameters and findings that can be used in our computational modeling.

Behavioral experiments: Many among the behavioral experiments dealing with audiovisual interactions focus on audiovisual integration, a well-studied manifestation of cross-modal interaction. Most of them try to provide insight on how, when, and where auditory and visual information are combined. Several works demonstrate the strong influence of audio on the perception of visual information [32]. The authors of [33], examine audiovisual simultaneity judgments: It is reported that after exposure to a fixed audiovisual time lag for several minutes, experiments on humans show shifts in their subjective simultaneity responses towards that lag. A related finding described in [34,35] is the brain's capability to rapidly recalibrate the presented audiovisual asynchrony, even when exposed to a single brief asynchrony.

The previously mentioned "pip and pop" effect [7] and other similar visual search experiments [36–38] are manifestations of the so-called temporal ventriloquism effect [32,39], which is an example of strong audiovisual integration. It is defined as the shift of a visual stimulus' onset and duration by a slightly asynchronous auditory stimulus, or as the "capture" of auditory time onsets over corresponding visual time onsets. A review on temporal ventriloquism is presented in [40], where the effects and the after-effects were studied, as well as the spatio-temporal criteria for stimuli binding. It has been found that the temporal ventriloquism effect is affected by temporal windows but is hardly affected by spatial discordance. A useful finding concerning the temporal windows is that audiovisual asynchrony cannot exceed 200 ms. The same finding appears also in [32], where it has also been found that audio influences visual timing perception even when sound trails the appearance of visual stimuli. In the same work, it is underlined that audio influences dynamic visual features and not the spatial ones.

An interesting theory is that sensory integration follows Bayesian laws [41,42]. The authors of [43] are based on the bayesian modeling of integration and they extend [41]. They also study the optimal time window of visual–auditory integration in relation to reaction time. They demonstrate that the time window acts as a filter determining whether information delivered from different sensory organs is registered close enough in time to trigger multi-sensory integration. In [44], behavioral experiments carried out in order to measure the temporal window of integration in audiovisual speech perception, also indicate a 200 ms window of integration.

These findings and conclusions will be used in order to fuse, in a behaviorally-inspired way, existing visual and auditory saliency models. The literature contains a rich set of visual saliency models, and a smaller set of auditory ones. Regarding the former, we focus on spatio-temporal ones, due to the need for a temporal component (more details in Section 3):

Visual saliency models: The authors of [45] present a review of the state-of-the-art in visual attention modeling, referencing about 65 different models and relative comparisons. In most cases, spatio-temporal models are an extension of spatial methods by incorporating dynamic features. We provide a brief overview of the various types: the biologically-inspired, the information theoretic, and the frequency/phase-selective ones.

Two seminal works [8,46] were the basis of many biologically-inspired attention models [47–50]. Itti et al. [9] provided an implementation of a bottom-up computational model for spatial visual saliency using three feature channels: intensity, color, and orientation, that was later extended into a spatio-temporal model for predicting saliency in video streams by the use of two additional features: motion and flicker [51]. There are other biologically-inspired models based on [9], either spatial [52–55] or spatio-temporal [56–59].

Most of the information-theoretic models are based on a Bayesian framework [60–63]. Zhang et al. [64] proposed a general framework for "Saliency Using Natural" (SUN) scene statistics, later to a spatio-temporal model [65]. Other works have exploited information-theoretic measures, like entropy, self/mutual information for spatial-only [66–72] or spatio-temporal models [68,71,73,74].

Another class of approaches estimates saliency in the frequency domain by frequency- or phase-selective tuning of the saliency map [75–77]. Some models are based on Fourier or discrete cosine transforms [75,78], while quaternion Fourier transform has also been employed for combining color, intensity, and motion features [77,79,80].

In [51,55,61], differences between the spatial orientation maps are employed as temporal features for saliency detection in videos. In [71], the authors extended their self-resemblance method by employing 3D local steering kernels for action and saliency detection in videos. In [81], a spatio-temporal filtering using temporal weighted summation is proposed for abnormal motion selection in crowed scenes, while in [82] researchers combine camera motion information with static features to study the differences between static and dynamic saliency

in videos. In [83], a perceptually based spatio-temporal computational framework for visual saliency estimation is presented, that produces both spatio-temporal and static energy volumes by using the same multi-scale filterbank based on quadrature Gabor filters in three dimensions (space and time). Also, in [84], a bottom-up saliency model based on the human visual system structure has been proposed.

From another point of view, based on learning, deep networks have been successfully applied for visual saliency: In [85], features from different network layers are used to train SVMs for fixated and non-fixated regions. Other approaches employ adaptation of pretrained CNN models for visual recognition tasks [86], while in [87], both shallow and deep CNN are trained end-to-end for saliency prediction. In [88], multiscale CNN networks are trained by optimizing common saliency evaluation metrics, while in [89], the authors extract fixation and non-fixation image regions to train end-to-end binary multiresolution CNN. The work of [90] shows that losses based on probability distance measures are more suitable for saliency rather than standard loss functions for regression. In [91], generative adversarial networks (GAN) are employed in order to better train end-to-end networks for fixation prediction. In [92], the authors proposed a two-stream CNN network based on RGB images and optical flow maps for dynamic saliency prediction. In [93], gaze transitions are learned from RGB, optical flow and depth information.

Auditory saliency models: As mentioned earlier, auditory saliency modeling has been investigated much less, and has initially been inspired by visual saliency modeling. One of the first biologically-inspired auditory saliency models has been proposed by Kayser et al. [10]. The auditory stimulus is converted into a time–frequency representation which is a sound spectrogram and yields an "intensity image", which serves as the model input. The output is a saliency map, which depicts how auditory saliency evolves across time and frequencies. In this context, it is structurally identical to Itti et al. visual saliency model [9, 51], but has a different interpretation, as it integrates the concept of time. The extracted features are the intensity, temporal contrast, and frequency contrast, in various scales (inspired by the function of auditory neurons). Each feature is extracted with filters modeling findings from auditory physiology: intensity filter corresponds to receptive fields with only an excitatory phase, frequency contrast filters to receptive fields with an excitatory phase and simultaneous side band inhibition, and temporal contrast ones to fields with an excitatory phase and a subsequent inhibitory one. These filters correspond to Gabor filters with suitable orientations. The model's output is a 2-D saliency map produced by summing the individual feature maps.

In [94], a model exploring the space of auditory saliency spanning pitch, intensity, and timbre is presented. It is based on the hypothesis that perception tracks the evolution of sound events in a multi-dimensional feature space and flags any deviation from background statistics as salient. Predictive coding corresponds to minimizing error between bottom-up sensations and top-down predictions. Corresponding mismatches signal the detection of a deviant, namely a salient event.

In [95], the authors propose a biologically-plausible auditory saliency model based on [10], augmented by orientation and pitch feature computation. The various features are integrated into a single 2-D saliency map using a biologically-inspired nonlinear local normalization algorithm, adapted from [96].

In [97], Bayesian surprise is applied to detect salient acoustic events. Kullback–Leibler divergence of the posterior and prior distribution is used as a measure of how "unexpected" and surprising newly observed audio samples are. This way, unexpected and surprising acoustic events are efficiently detected.

In the context of acoustic salient event detection, the model proposed in [14] measures Dominant Teager energies over a 1D Gabor filterbank applied on the audio signal. In [98], the authors examine whether saliency scores are modified just after auditory salient events. They develop two different auditory saliency models, the discrete energy separation algorithm (DESA) and the energy model that provide saliency curve as an output. The most salient auditory events are extracted by thresholding these curves and the authors examine some eye movement parameters just after these events concluding that audio impact on visual saliency is not reinforced specifically after salient auditory events.

3. Computational audiovisual saliency modeling

The main focus of this work is to fuse, in a behaviorally-inspired way, individual auditory and visual saliency models in order to form a 2-D audiovisual saliency model and investigate its plausibility. We essentially try to combine several theoretical and experimental findings from neuroscience with signal processing techniques. A high-level overview of the model is presented in Fig. 2. An auditory and a visual stimulus serve as input to an auditory and a visual spatio-temporal saliency model where saliency features are computed. At some point, that will be described later in this section, the two saliencies are appropriately fused in order to form an audiovisual saliency map. The majority of our parameters and fusion schemes which are discussed below, are inspired by cognitive research, and findings from behavioral experiments.

3.1. From auditory saliency map to auditory saliency curve

Most of the existing auditory saliency models yield a 2-D saliency map as output, as described in the previous section. Usually, this map is a time–frequency representation of auditory saliency. However, in this work we are more interested in the evolution of auditory saliency through time, rather than how it is distributed among the involved frequencies. Also, since for a visual input we obtain a 2-D saliency map, it seems more intuitive for an auditory input (which is 1-D and non-spatial) to obtain an 1-D saliency curve. Thus, if the auditory output is a saliency map, we have to appropriately process it to obtain an 1-D saliency curve. The same reasoning has also been followed in the past, both in [95], where the time curve was obtained by adding saliency values across frequencies, and in [99], where it was obtained by maximizing over frequencies. The latter approach appears also in [94], where the final temporal saliency score is the maximum for each time instance, but it has additionally been behaviorally-validated for capturing salient events in [10]. Therefore, we follow the same approach. With $M_a(\ell, f)$ we denote the auditory saliency map that is a function of time ℓ and frequency f and with $S_a(\ell)$ the auditory saliency curve, computed as:

$$S_a(\ell) = \max_f M_a(\ell, f) \tag{1}$$

3.2. Audiovisual temporal window of integration

As briefly stated in Section 2, behavioral experiments indicate that synchrony between an auditory and a visual stimulus (e.g. a slamming door) results in a strong audiovisual integration. However, they also indicate that partially asynchronous stimuli can still result in audiovisual integration, i.e., audiovisual integration can be tolerant to an amount of asynchrony. Most related works agree to an approximately 200ms long maximum temporal window of audiovisual integration [7,40,44].

In order to incorporate this finding in our computational model, instead of taking into account only the present values of auditory and visual saliencies, we appropriately filter the auditory saliency curve: As presented in Section 2, audition dominates vision in temporal tasks [100–102], and it influences vision even when preceding or trailing it [32]. These facts indicate that we should take into account not only current auditory saliency values, but properly weigh past and future values as well. A suitable filter should favor the synchronized stimuli, by weighting higher the present auditory saliency value, but also include past and future values with attenuation. Thus, we employ a Hanning window on the auditory saliency curve, with 200 ms length

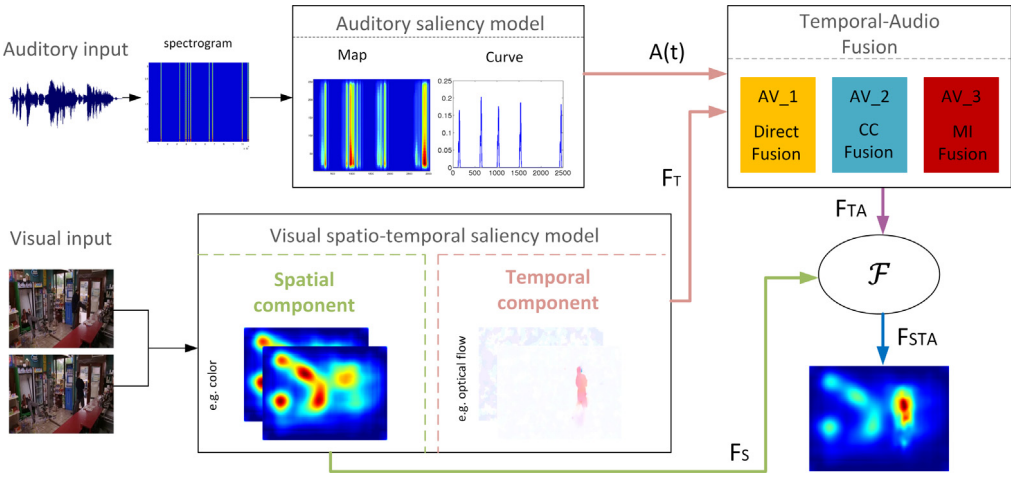

Fig. 2. An overview of the 2D audiovisual saliency model (better viewed in color). Auditory and visual streams constitute the inputs to the auditory and spatio-temporal visual saliency models respectively. These streams are individually processed for saliency extraction. Auditory saliency is fused with the temporal visual saliency map with one out of three available fusion schemes. Lastly, the spatial visual map is also integrated according to the visual model's initial fusion methodology. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and center it on the current time instance (other similar windows could also be employed without significant differences). After windowing, we apply a moving average, thus obtaining a new saliency curve. Thus, the final saliency curve $A(t)$ is computed as:

$$A(t) = \frac{1}{2N+1} \sum_{\ell=-N}^{N} S_a(t+\ell) H(\ell) \tag{2}$$

where t is the video time index, ℓ the audio sample index, and $H(\ell)$ the Hanning window with $2N+1$ window length.

3.3. Audiovisual saliency fusion

Since we aspire to combine auditory and visual saliency in order to obtain an audiovisual saliency model, the most important issue to be addressed is where and how fusion will take place. Auditory and visual saliency representations are inherently non-comparable modalities with different dimensions and ranges. Our proposed fusion schemes hypothesize that due to the dynamic nature of the audio features, they influence only the temporal/dynamic visual features, and not the spatial ones [32]. Our hypothesis is not arbitrary, as there is evidence for this influence in the literature [101,103]. In some works, interactions of audio with specific dynamic visual features are investigated, such as flicker in [100,104], and motion in [105]. Thus for our model, audiovisual fusion is performed between auditory saliency and temporal visual saliency.

Another equally important issue is how the audiovisual fusion will be performed, since the two modalities are non comparable. In the absence of audio, temporal visual saliency maps should be left unaltered, while when present, its saliency should weigh them appropriately. We have experimented with three different fusion schemes, inspired by well-known techniques for combining different modalities. Fusion is applied between auditory saliency and each individual temporal feature of visual saliency separately. This fusion results in a joint temporal-audio map F_{TA}, where the audio influence has been integrated into the 2-D temporal visual saliency map. After temporal-audio fusion, the spatial visual component is also integrated appropriately, according to each method's fusion strategy $\mathcal{F}$ (for visual-only saliency), thus resulting in the final spatio-temporal-audio saliency map, denoted by F_{STA}, as also depicted in the lowest right corner of Fig. 2:

$$F_{STA} = \mathcal{F}(F_S, F_{TA}) \tag{3}$$

where F_S is the spatial saliency map. We focus on the temporal-audio fusion, because the final fusion $\mathcal{F}$ is dependent on the specific spatio-temporal visual saliency model that is employed. For example, in Itti et al. model [9], $\mathcal{F}$ is an averaging of the individual saliency maps. The next sections describe the proposed fusion schemes between auditory and temporal visual saliency, in order to compute F_{TA}.

3.3.1. Direct fusion of saliencies (*direct fusion*)

We experiment with fusing audio saliency curve with the dynamic visual saliency map directly, in a simple multiplicative manner, separately for each temporal visual feature:

$$F_{TA}(x,y,t) = F_T(x,y,t)(1 + A(t)) \tag{4}$$

where x,y are the pixel coordinates, F_{TA} is the fused map, F_T represents a single temporal/dynamic feature map of visual saliency and A is the auditory saliency curve. This audiovisual fusion scheme appears in [21], but for point-wise multiplication between visual and spatial audio maps, which have the same dimensions, since the spatial auditory map is the source location map. In our case, the auditory saliency value weighs uniformly all temporal visual saliency pixel values.

3.3.2. Cross-correlation between audio and video as weight (CC fusion)

In [107], cross-correlation is proposed as a measure of audiovisual integration. Cross-correlation between multiple sensory signals is an important cue for causal inference: signals originating from a single event normally share a tight temporal relation, due to their dependence on the same underlying event. Conversely, when multiple signals are generated by independent physical events, their temporal structures are normally unrelated. Specifically, signals with a similar fine-temporal structure, and, thus, a high cross-correlation, are more likely inferred to originate from a single underlying event and hence will be integrated more strongly.

Here, according to the temporal window of integration modeling, cross-correlation with a restricted lag τ should be used in order to fuse visual and auditory saliencies, because, assuming that audio and video originate from the same event, they can be perceived as such, if their asynchrony does not exceed 200 ms. Thus, cross-correlation lag cannot exceed 200 ms. From this point on we denote cross-correlation by R_{TA}. We compute cross-correlation for time windows of the audio and visual saliencies. More specifically, we choose a time window of 1 second (and $2k+1$ number of frames) and we compute cross-correlations between the time series of audio and the time series of every pixel of temporal saliency feature maps. Subsequently, the max cross-correlation value weighs the current pixel's value:

$$F_{TA}(x,y,t) = F_T(x,y,t)(1 + \max_{\tau} R_{TA}(x,y,t,\tau)),$$

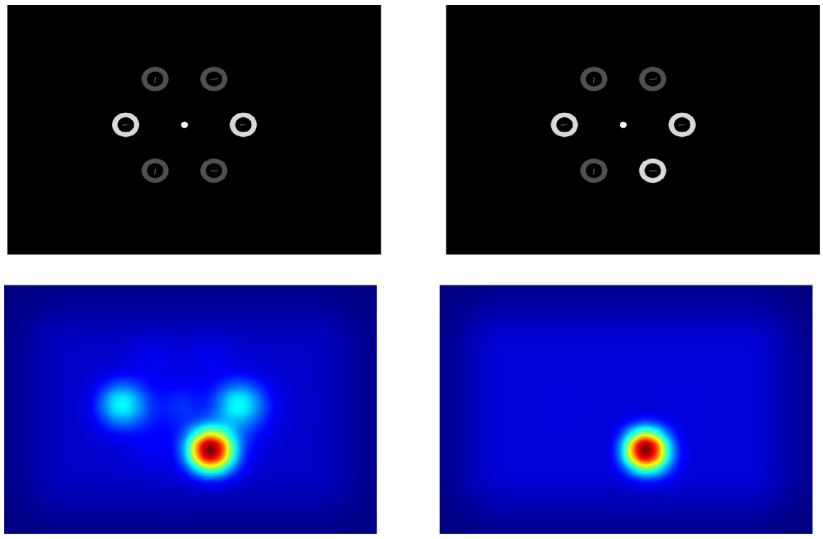

Fig. 3. These two figures from [106] depict the "sine vs square" stimuli for the square modulation case, during a target flicker (the vertical line in the lower right corner). Below are the visual (left) and audiovisual saliency map (right). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$t - n_{cc} \leq \tau \leq t + n_{cc} \tag{5}$$

with

$$R_{TA}(x, y, t, \tau) = \frac{1}{2k+1} \sum_{m=t-k}^{t+k} F_T(x, y, m) A(m - \tau),$$
$$t - n_{cc} \leq \tau \leq t + n_{cc} \tag{6}$$

where n_{cc} denotes the possible values of the lag τ (it cannot exceed 200 ms) and R_{AT} the cross-correlation between audio and temporal visual saliencies.

3.3.3. Mutual information between audio and video as weight (MI fusion)

Inspired by [108,109], we examine mutual information between audio and visual saliencies as an expression of audiovisual simultaneity of an event. We assume that audio and visual saliencies come from a joint probabilistic process, which is stationary and Gaussian in a short period of time [108]. If we denote with $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ this joint Gaussian distribution, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be estimated from the audiovisual data per frame. For a specific frame t:

$$\boldsymbol{\mu}(x, y, t) = b \begin{bmatrix} A(t) \\ F_T(x, y, t) \end{bmatrix} + (1 - b)\boldsymbol{\mu}(x, y, t - 1) \tag{7}$$

$$\boldsymbol{\Sigma}(x, y, t) = \frac{1}{1+a} \left(a \left(\begin{bmatrix} A(t) \\ F_T(x, y, t) \end{bmatrix} - \boldsymbol{\mu}(x, y, t - 1) \right) \left(\begin{bmatrix} A(t) \\ F_T(x, y, t) \end{bmatrix} - \boldsymbol{\mu}(x, y, t - 1) \right)^T + \boldsymbol{\Sigma}(x, y, t - 1) \right) \tag{8}$$

where a, b are pre-defined weights within $[0, 1]$, that control the dependence of the current values on the past ones. Mutual information between audio and video is then computed as:

$$I(x, y, t) = -\frac{1}{2} \log \left(\frac{|\boldsymbol{\Sigma}_A(t)||\boldsymbol{\Sigma}_{F_T}(x, y, t)|}{|\boldsymbol{\Sigma}(x, y, t)|} \right) \tag{9}$$

and $\boldsymbol{\Sigma}$ can be expressed as:

$$\boldsymbol{\Sigma}(x, y, t) = \begin{bmatrix} \boldsymbol{\Sigma}_A(t) & \boldsymbol{\Sigma}_{AF_T}(x, y, t) \\ \boldsymbol{\Sigma}_{AF_T}(x, y, t)^T & \boldsymbol{\Sigma}_{F_T}(x, y, t) \end{bmatrix} \tag{10}$$

In the case of one audio and one visual feature, which is our case, (9) and (10) are simplified as:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_A(t)^2 & \sigma_{AF_T}(x, y, t) \\ \sigma_{AF_T}(x, y, t) & \sigma^2_{F_T}(x, y, t) \end{bmatrix} \tag{11}$$

$$I(x, y, t) = -\frac{1}{2} \log \left(1 - \rho^2(x, y, t)\right) \tag{12}$$

$$\rho(x, y, t) = \frac{\sigma_{AF_T}(x, y, t)}{\sqrt{\sigma_A(t)\sigma_{F_T}(x, y, t)}} \tag{13}$$

where σ_{AF_T}, σ_A, and σ_{F_T} are the scalar estimates of audio-visual feature covariance and the variances of the audio-only and visual-only faetures respectively, and ρ is the Pearson's Correlation Coefficient. The fused map is computed as:

$$F_{TA}(x, y, t) = F_T(x, y, t)(1 + I(x, y, t)) \tag{14}$$

4. Evaluation

4.1. Evaluation metrics

Since we are addressing a fixation prediction problem, which is primarily a visual task where the auditory influence has been incorporated into a visual saliency map, the evaluation metrics we adopt consist of widely used visual saliency evaluation metrics [45,110].

We denote the output of our model by Estimated Saliency Map (ESM). In eye-tracking experiments, the Ground-truth Saliency Map (GSM) is the map built from eye movement data. In behavioral experiments, inspired by [6], and due to the lack of eye-tracking data, GSM consists of the ground truth target location in the sense that only the target is salient. The employed metrics are the following [45,110,111]:

(1) Linear Correlation Coefficient (CC): It measures the strength of a linear relationship between the continuous GSM[1] and ESM. When CC is close to $+1/-1$ there is almost a perfect linear relationship between the two variables.

(2) Normalized Scanpath Saliency (NSS): For an ESM normalized to zero mean and unit standard deviation, NSS is the average of the response values on ESM at human eye positions. It shows how many times over the whole ESM's average is the ESM value at each human fixation. The final NSS value is the mean over all viewers fixations.

(3) Area Under Curve shuffled (AUCs): For eye-tracking experiments, *shuffled AUC* is employed, according to [64,110], where negative set is formed by sampling fixation points from 10 random frames. Since for computing AUC a positive and a negative set are needed, for behavioral

[1] In this case the continuous GSM map arises by convolving the binary fixation map with a gaussian kernel of size 10 for video dimension of 640 × 480.

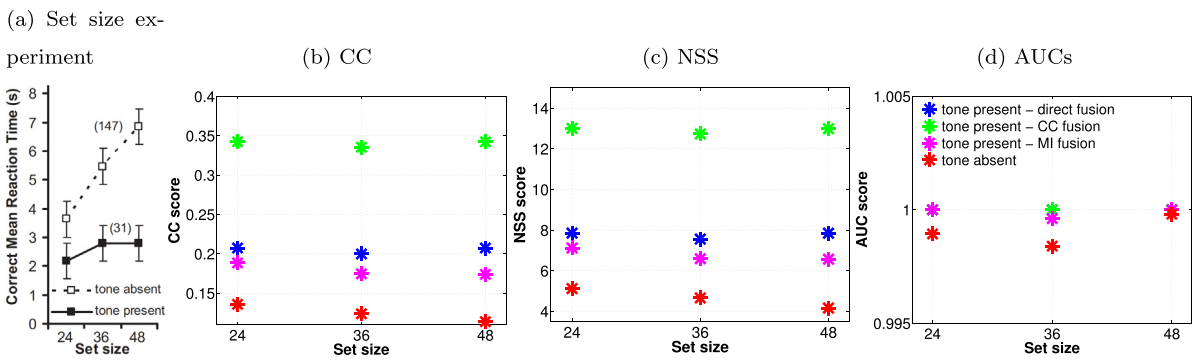

Fig. 4. (a) Original figure from [7] and (b, c, d) *CC, NSS, AUCs* for the set size experiment with all fusion schemes. Blue color denotes direct fusion, green color denotes CC fusion, magenta denotes MI fusion, and red color the results when tone is absent. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

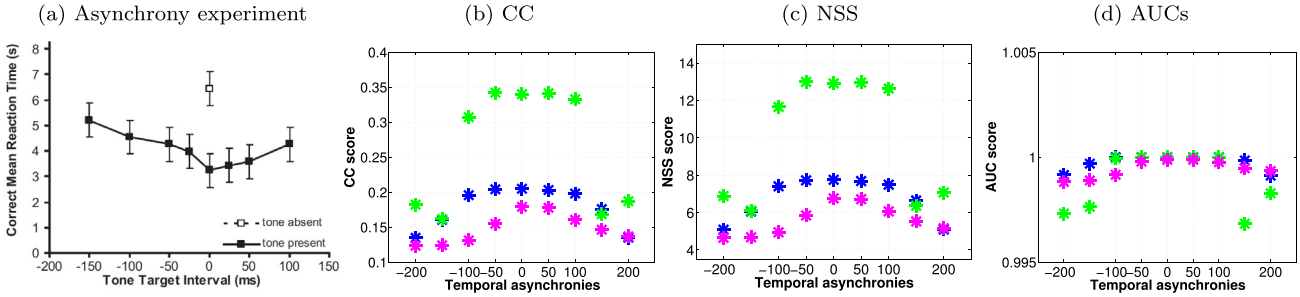

Fig. 5. (a) Original figure from [7] and (b, c, d) *CC, NSS, AUCs* for the temporal asynchrony experiment. Minus offsets refer to audio stream preceding the visual one. Different colors represent the same fusion schemes as in the above figure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

experiments, the former consists of the target location while the latter of a subset of points sampled from the distractors' positions. With the ESM as a binary classifier between the two sets, a ROC curve is formed by thresholding over ESM, plotting true positive vs. false positive rate. *AUCs* is then the area underneath the average of all ROC curves. A perfect prediction implies an $AUCs = 1$.

4.2. Behavioral experiments

The first part of the evaluation strategy involves comparison with results from behavioral experiments that have investigated aspects of audiovisual integration. One such category are audiovisual stimuli from visual search tasks. In such behavioral experiments the users' task is to detect a target among some distractors without scanning the whole image, but instead by focusing on the center of the screen. Their performance is measured by their Response Time (RT), which signifies the time that elapses between the target appearance and its detection by the user (usually by pressing a button). This evaluation aims to assess whether our model reproduces findings from human behavioral experiments using the concept of saliency instead of RT, in the sense that a more salient target needs less time to be detected and vice-versa [46,112,113]. The evaluation metrics employed, thus, are saliency metrics, the ones which have been described above.

Regarding saliency models, the biologically-inspired Itti et al. model [9,51] is employed for visual saliency, which has been already validated with human experiments [6], while for auditory saliency, the biologically-inspired Kayser et al. [10] model is used. More specifically, for Itti et al. model [9,51], the spatial component comprises of color, orientation, and intensity features, while the temporal one comprises of flicker and motion. Audiovisual fusion is performed separately for flicker and motion, with three different choices for fusion, as previously described.

Regarding the employed stimuli, as discussed before, they are stimuli used in visual search tasks and particularly the "pip and pop" and "sine vs. square" stimuli [7,106]. The former, depicted in Fig. 1, have already been presented briefly in Section 2. The visual and the audiovisual saliency maps for two example successive frames are also depicted in the same figure. Regarding the "sine vs. square" stimuli, they are straight lines as well, surrounded by annuli whose luminance changes continuously with time in gray scale, following a sine or square modulation. The target's luminance changes are either synchronized with a non-spatial audio pip in phase or with a 180° phase difference (square or sine modulated) or there is no audio. The target is a horizontal or vertical line and distractors may have all other orientations. This experiment is a comparative one: The authors claim that audiovisual integration requires transient events and they compare the same audiovisual stimulus with two different modulations, the sine (gradual) and the square (transient). An example of these stimuli can be found in Fig. 3, where two successive frames from a square modulation case are depicted.

For the following experiments and results, first the actual behavioral experiment and its corresponding findings are described, and subsequently we present and discuss the results produced by our model for the same inputs.

4.2.1. "Pip and pop" set size experiment

In [7], experiments carried out by the authors indicate that for the "pip and pop" audiovisual stimulus case (visual target color change with synchronized audio pip), RT is independent of the number of distractors (i.e. set size), while, for the visual-only stimulus, RT changes analogously with the set size (increases for larger set size), probably because there is no integration and a serial search is required. These findings are depicted in Fig. 4(a) for three different set sizes, which is an original figure from [7]. Using the same visual and audiovisual "pip and pop" stimuli as input, we investigate if our model reproduces the same finding expressed in terms of saliency. In Fig. 4, we present our results for *CC, NSS* and *AUCs* metrics.

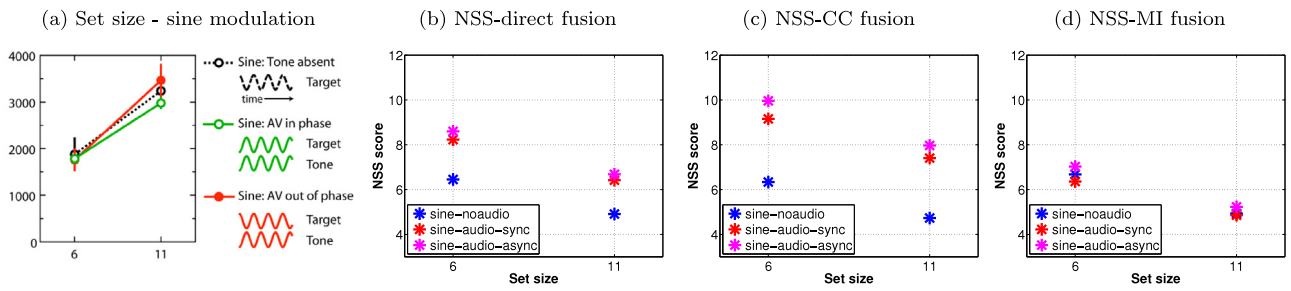

Fig. 6. (a) Original figure from [106] regarding sine modulation and (b, c, d) the *NSS* results for the fusion schemes, for the set size experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

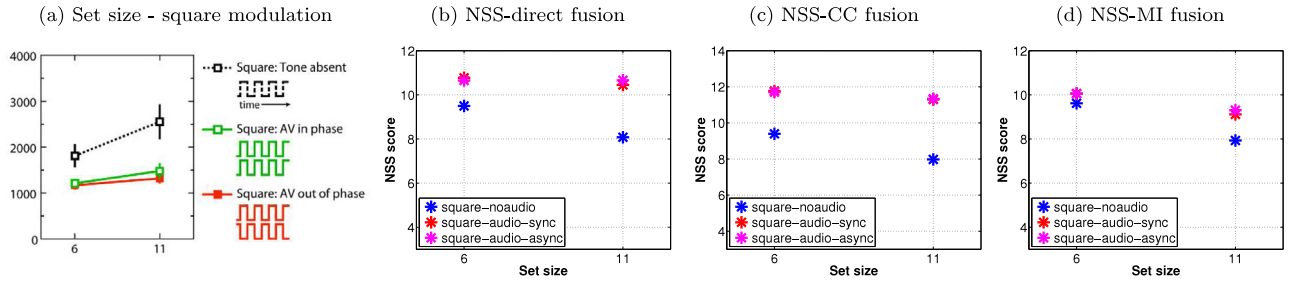

Fig. 7. (a) Original figure from [106] regarding square modulation and (b, c, d) the *NSS* results for the fusion schemes, for the set size experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For the *CC* and *NSS* metrics, although the resulting slopes of the curves are not the same with the original figure, the behavior is captured well enough: saliency slightly decreases for a larger set size in the visual case, while it remains almost constant for the audiovisual case, for all fusion schemes. Generally, saliency is higher for the audiovisual case, which is also in line with behavioral results. Regarding *AUC*, we notice that it yields an almost perfect prediction for both cases, thus, maybe due to the nature of these stimuli, it probably cannot capture well the differences between the visual and the audiovisual case.

4.2.2. *"Pip and pop" temporal asynchrony experiment*

A second behavioral experiment from [7] investigates audiovisual integration in terms of asynchrony tolerance, namely when audio and visual segments that belong to the same event are asynchronous to each other. The findings indicate that audiovisual integration can tolerate a certain amount of asynchrony. The authors depict how asynchrony is related to RTs, showing that the larger the asynchrony is, the more the performance drops and RT increases. Also, they claim that for the same amount of asynchrony, when the auditory stream trails the visual one, RT decreases more (saliency increases) than in the opposite case. All the above are depicted in Fig. 5(a), an original figure from [7]. Using as input the stimuli employed in the behavioral experiment, we compute the saliency results from our model, depicted in Fig. 5.

Here, *CC* and *NSS* seem to reproduce well enough the behavioral observations, yielding the maximum saliency when the auditory and the visual streams are synchronized, and decreasing gradually as the amount of asynchrony increases. Also, when audio stream trails the visual one, saliency is slightly higher than in the opposite case, as observed in the behavioral experiments as well. Again, *AUC* does not reproduce equally well the corresponding behavioral results. Among the several fusion schemes, the best results are given by the *MI fusion* scheme, where the curve exhibits exactly the same behavior with the original one. The *CC fusion* scheme does not capture exactly the form of the behavioral results.

4.2.3. *"Sine vs. square" set size experiment*

In a similar fashion to the first "pip and pop" experiment, in [106] experiments indicate that for visual-only stimuli, RTs increase analogously with the set size (target saliency decreases). The same effect appears even when audio is present, if luminance and audio change with sine modulation (gradually). The authors attribute this effect to the lack of audiovisual integration. On the contrary, when a brief synchronized audio pip accompanies the target color flicker and both are square-modulated, RT is independent of the set size. The same effect appears even when the audio pip has a 180-phase difference with luminance modulation. The original figures from [106] presenting these results are Figs. 6(a) and 7(a): We investigate whether our model does exhibit the same behavior. In Fig. 6, we present our results for the *NSS* metric (due to lack of space, we do not present *CC*, which was very similar) for the sine modulation and in Fig. 7 the corresponding results for the square one.

We observe that for both sine and square modulation, our results indicate a similar behavior to the behavioral ones. When there is no audio pip, saliency decreases when set size increases in all cases. The same happens for sine modulation, whether the pip is synchronized or 180-desynchronized with the luminance change. On the contrary, for the square modulation, we can notice that saliency remains high and almost constant independently of the set size, for both synchronized and 180-desynchronized audiovisual stimuli, exactly as depicted in the original paper figure. *Direct fusion* scheme yields better results than *CC* and *MI* regarding square modulation.

4.3. *Eye-tracking data collection on SumMe and ETMD databases*

For the purposes of experimental evaluation with eye-tracking data, since there are only a few databases with audiovisual eye-tracking data, we decided to collect such data for two databases, SumMe [114] and ETMD [83]. The SumMe database contains 25 unstructured videos, while the ETMD contains 12 videos from six different hollywood movies, both summing up to 37 videos totaling approximately 2 h and 171,000 frames. For this reason, the group of participants and of videos were split into two equivalent groups containing the half number of people and videos, respectively. Thus, each video was seen by 10 different subjects. The subjects were recruited through the National Technical University of Athens, with ages ranging from 23 – 55 (mean 35). Almost all subjects were naive as to the purposes of the experiment and they all had normal vision. The employed videos ranged from 38

to 388 s in length and they were converted from their original sources to a MOV video format.

Eye movements were binocularly monitored via a SR Research Eyelink 2000 desktop mounted eye-tracker with 1000 Hz sampling rate. Videos were displayed on a 1600 × 900 monitor at a 90 cm distance from the viewer. Audio was delivered in stereo, through headphones. A chin and headrest was used during the experiment, in order to ensure the viewer's minimal movement and avoid continuous calibration. Presentation was controlled using the SR Research Experiment Builder software. The subjects that participated in the experiment were informed only that they would watch some videos and that they should avoid moving during a video playback. The order of the clips was randomized across participants. The whole experimental procedure for each participant was approximately 90 min long, including instructions, calibration, testing, and short breaks if needed.

Regarding calibration, a 13-point binocular calibration preceded the experiment. Before each video, if central fixation accuracy was exceeding a pre-defined threshold of 0.5°, a full calibration was repeated. The central fixation marker also served as a cue for the participant and offered an optional break-point in the procedure. After checking for a central fixation, the start of each trial was manually triggered. Regarding post-processing, the 1000-Hz raw eye-tracking recordings were sampled down to match each video's frame rate. One sample frame per video with its corresponding eye-tracking data superimposed, and the distribution of eye-tracking data for the whole video can be found in Figs. 8 and 9 for SumMe and ETMD databases for all videos. The data are publicly released and can be found in http://cvsp.cs.ntua.gr/research/aveyetracking.

4.4. Eye-tracking experiments

We evaluate four different visual spatio-temporal models fused with audio via the three different fusion schemes and Min et al. [11] audiovisual saliency model, with SR [75] for static model (which is to the best of our knowledge the only publicly available audiovisual model) on 6 eye-tracking databases: DIEM [115], AVAD [11], Coutrot1 and Coutrot2 [27,28], SumMe [114], and ETMD [83,116] that present rather complicated and challenging stimuli.

For the evaluation of the fusion schemes, we experiment with several state-of-the-art publicly available visual spatio-temporal saliency models. We choose one model from each one of the basic approaches in visual saliency: a biologically-inspired model, Itti et al. [9,51], described previously, one information theoretic, SDSR [71], a frequency domain one, PQFT [77], and a developed baseline deep learning framework based on [87]. SDSR model includes one spatial and one temporal model, thus to obtain a visual spatio-temporal model, we simply add the final spatial and temporal saliency maps. Regarding the PQFT model, from the image's quaternion representation, we employ motion as the temporal component and two color and one intensity channels as the static part. Then we calculate the spatio-temporal saliency map by applying the Quaternion Fourier Transform as in [77].

Regarding the deep model, a hybrid approach that incorporates a state-of-the-art CNN network for static saliency and an optical flow estimation for temporal saliency has been employed. For the static component, we used the publicly available deep model from [87] (pre-trained only on static images). For temporal visual saliency, we extracted warped optical flow maps according to [117], which is based on the TVL1 optical flow algorithm [118]. Temporal moving averaging is applied over ten successive frames to smooth and remove the noise from optical flow estimation in x and y directions. Finally, we apply Difference-of-Gaussians (DoG) filtering to the optical flow magnitude [119]. For the final spatio-temporal saliency map, we add the two normalized maps. For auditory saliency, the biologically-inspired Kayser et al. [10] model is used.

We do not aspire to fully optimize our models to achieve the highest possible results compared to the literature, but to assess if visual saliency performance is improved when fused with auditory saliency. Thus, it is rather a comparative study between visual and audiovisual combinations. We also evaluate the Min et al. [11] audiovisual attention model with SR for static model, as released by the authors.

Table 1
Results for the various models and fusion schemes (with acronym and brief description) on DIEM database.

	Model/Fusion scheme	CC	NSS	AUCs
Itti_V	[51]/-	0.195	1.121	0.507
Itti_AV1	[51]/Direct	0.199	1.150	0.588
Itti_AV2	[51]/CC	0.196	1.127	0.521
Itti_AV3	[51]/MI	0.196	1.128	0.511
PQFT_V	[77]/-	0.119	0.725	0.555
PQFT_AV1	[77]/Direct	0.118	0.713	0.554
PQFT_AV2	[77]/CC	0.119	0.717	0.552
PQFT_AV3	[77]/MI	0.118	0.711	0.554
SDSR_V	[71]/-	0.088	0.524	0.542
SDSR_AV1	[71]/Direct	0.089	0.528	0.542
SDSR_AV2	[71]/CC	0.108	0.642	0.556
SDSR_AV3	[71]/MI	0.108	0.645	0.559
Deep_V	Modif.[87]/-	0.265	1.563	0.636
Deep_AV1	Modif.[87]/Direct	**0.270**	**1.585**	**0.636**
Deep_AV2	Modif.[87]/CC	0.243	1.4192	0.611
Deep_AV3	Modif.[87]/MI	0.258	1.518	0.633
Min et al. (SR)	[11,31]	0.121	0.722	0.593

4.4.1. DIEM database

DIEM database [115] consists of 84 movies of all sorts, sourced from publicly accessible repositories, including advertisements, documentaries, game trailers, movie trailers, music videos, news clips, and time-lapse footage. Thus, the majority of DIEM videos are documentary-like, which means that audio and visual information do not correspond to the same event. Eye movement data from 42 participants were recorded via an Eyelink eye-tracker, while watching the videos in random order and with the audio on. We evaluate our models on this database and we compare the performance of visual-only models with audiovisual ones. Results are depicted in Table 1. Regarding Itti et al. [51] and SDSR [77], the audiovisual models outperform the visual ones for all metrics and fusion schemes, while audiovisual deep models with *direct fusion* yield the highest saliency result.

4.4.2. AVAD database

Finally, we apply our models on AVAD database [11] that contains 45 short clips of 5-10 s duration with several audiovisual scenes, e.g. dancing, guitar playing, bird signing, etc. The majority of the videos might contain soundtrack or other ambient sound, but they also contain one dominant sound that corresponds to a visual event. Additionally, the joint audiovisual event is always present and usually centered throughout the whole video duration. Eye-tracking data from 16 participants have been recorded. For this database, we have re-evaluated the Min et al. (SR) [11] model with our evaluation framework. The results are presented in Table 2. Audiovisual models outperform the visual ones for all almost all combinations and metrics except for the deep models, where the visual models perform better than the audiovisual ones for the first two metrics. Deep models also outperform Min et al. [11] model, which achieves the second best performance in terms of CC and NSS. This might be due to the fact that deep models in general learn to capture well semantic information, like faces, musical instruments, etc., especially when those appear in the center of the image. These clips are very short and specific and contain the audiovisual event without any transition from visual to audiovisual. Thus, probably there is nothing more to be highlighted by the audio that has not been already captured by the visual deep models. For AUCs metric, that is more robust to center bias, there is still a slight improvement.

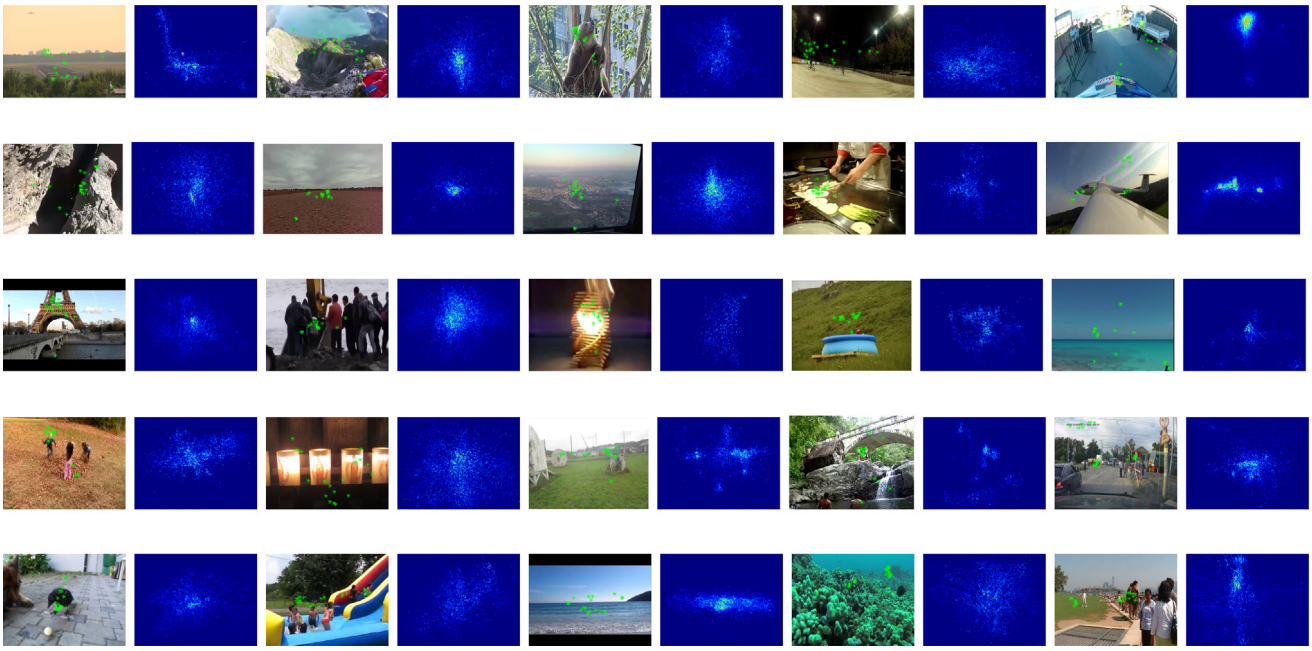

Fig. 8. Sample video frames with eye-tracking data from SumMe database, along with the distribution of eye-tracking data for the whole video.

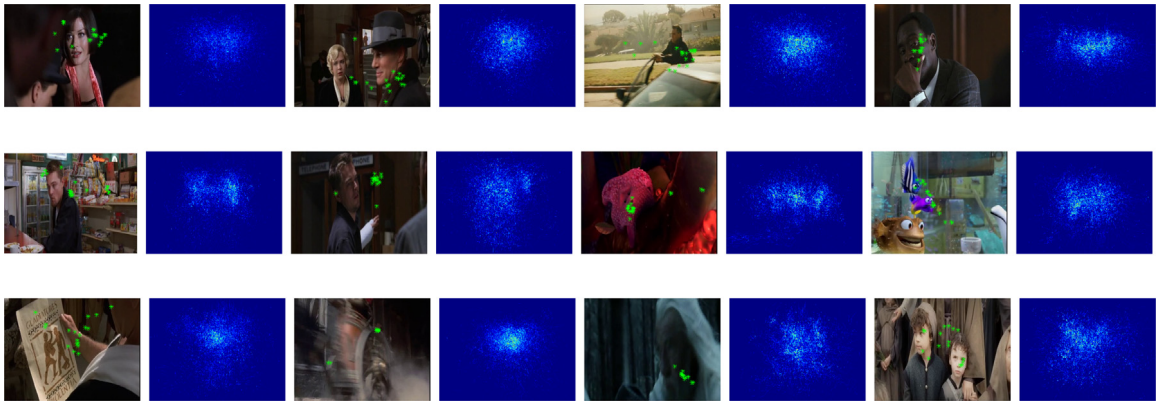

Fig. 9. Sample video frames with eye-tracking data from ETMD database, along with the distribution of eye-tracking data for the whole video.

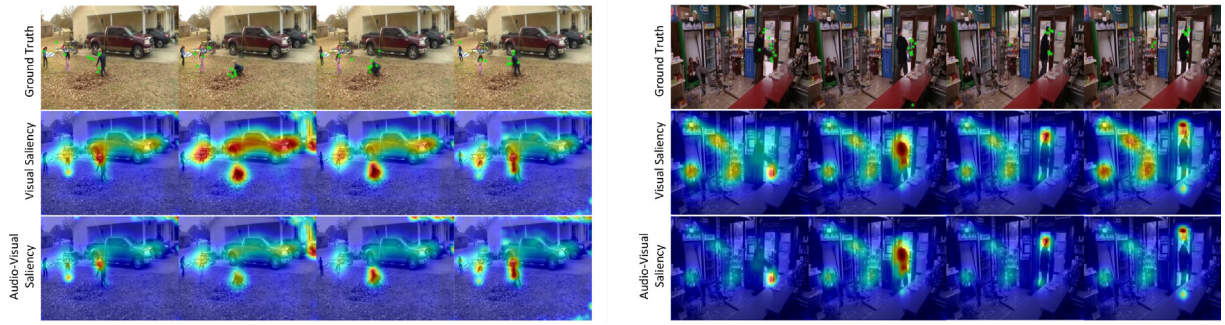

Fig. 10. Example of consecutive frames where the collected eye-tracking data have been overlaid, for SumMe (left) and ETMD (right). Second and third row depict the corresponding visual-only and audiovisual saliency maps.

4.4.3. Coutrot databases

We also apply our model on Coutrot databases [27,28]: Coutrot1 contains 60 clips with dynamic natural scenes split in 4 visual categories: one/several moving objects, landscapes, and faces. Eye-tracking data from 72 participants have been recorded. Coutrot2 contains 15 clips of 4 persons in a meeting and the corresponding eye-tracking data from 40 persons. The results are presented in Table 3. Contrary to the DIEM database, here the majority of the videos contains scenes where video and audio originate from the same event. Thus, we expect the results to be better than in DIEM for the audiovisual models. Indeed, audiovisual models outperform the visual ones for all combinations and

Table 2
Results for the various models and fusion schemes on AVAD database. (* Re-evaluated with the current evaluation framework.).

	CC	NSS	AUCs
Itti_V	0.154	1.436	0.529
Itti_AV1	0.172	1.627	0.542
Itti_AV2	0.161	1.502	0.532
Itti_AV3	0.154	1.434	0.528
PQFT_V	0.093	0.886	0.527
PQFT_AV1	0.095	0.911	0.527
PQFT_AV2	0.095	0.909	0.525
PQFT_AV3	0.096	0.920	0.528
SDSR_V	0.098	0.922	0.521
SDSR_AV1	0.099	0.929	0.522
SDSR_AV2	0.117	1.091	0.526
SDSR_AV3	0.115	1.072	0.526
Deep_V	**0.199**	**1.893**	0.552
Deep_AV1	0.194	1.844	**0.553**
Deep_AV2	0.192	1.830	0.551
Deep_AV3	0.196	1.859	0.552
Min et al. (SR)*	0.174	1.652	0.550

Table 3
The results for the various models and fusion schemes on Coutrot databases.

	Coutrot1			Coutrot2		
	CC	NSS	AUCs	CC	NSS	AUCs
Itti_V	0.181	1.015	0.544	0.164	1.362	0.593
Itti_AV1	0.183	1.062	0.559	0.239	2.005	0.632
Itti_AV2	0.187	1.055	0.548	0.177	1.475	0.600
Itti_AV3	0.182	1.013	0.543	0.166	1.373	0.593
PQFT_V	0.128	0.845	0.543	0.162	1.584	0.588
PQFT_AV1	0.130	0.853	0.544	0.166	1.615	0.588
PQFT_AV2	0.128	0.845	0.542	0.159	1.550	0.582
PQFT_AV3	0.131	0.862	0.544	0.169	1.640	0.588
SDSR_V	0.115	0.674	0.539	0.072	0.622	0.560
SDSR_AV1	0.116	0.677	0.540	0.073	0.627	0.561
SDSR_AV2	0.128	0.748	0.542	0.096	0.829	0.586
SDSR_AV3	0.128	0.744	0.544	0.098	0.839	0.589
Deep_V	0.266	1.585	0.586	0.248	2.094	0.624
Deep_AV1	**0.270**	**1.650**	**0.591**	**0.253**	**2.157**	**0.627**
Deep_AV2	0.264	1.598	0.588	0.249	2.113	0.623
Deep_AV3	0.266	1.595	0.589	0.251	2.132	0.626
Min et al. (SR)	0.115	0.666	0.550	0.127	1.043	0.605

almost all metrics. In both Coutrot1 and Coutrot2, the best results are achieved by the audiovisual deep models with *direct fusion*.

4.4.4. SumMe database

SumMe database [114] contains 25 unstructured videos as well as their corresponding multiple-human created summaries, which were acquired in a controlled psychological experiment. As mentioned before, we have collected eye-tracking data, and use them for evaluation. A few frames with their eye-tracking data and the corresponding visual and audiovisual saliency maps (using Deep models) are depicted on the left side of Fig. 10. The viewers attend mostly to the moving child instead of the car, which is mirrored better on the audiovisual maps compared to the visual ones. Results are presented in Table 4. The audiovisual combinations yield better results than the visual-only models for most cases: For Itti et al. [51] the best performance is achieved with *CC fusion*, while for SDSR and PQFT *direct* and *MI fusion* yield the best results respectively. Regarding deep models, only *direct fusion* yields a slightly improved *AUCs* compared to visual-only models, a fact that has also been recently observed in the comparison of spatial-only to spatio-temporal deep models for visual saliency [120].

4.4.5. ETMD database

ETMD database [83,116] contains 12 movie clips from 6 movies. Movie clips are complex stimuli because they are highly edited and contain a lot of semantics. A few frames with their eye-tracking data and the corresponding visual and audiovisual saliency maps (with Deep models) are depicted on the right side of Fig. 10. The viewers attend mostly to the slamming door, which is mirrored better on the audiovisual maps compared to the visual ones. The results appear in Table 4, and indicate a similar trend as in SumMe database. The audiovisual combinations yield better results than the visual-only models for almost all metrics except for deep models, where they are only comparable to the visual-only. This may be due to the fact that movies contain a lot of semantic information that is already integrated into the spatial-only model (during training on large image datasets). These results indicate that for proper audiovisual integration, top-down information is also required, an observation that also highlights the appropriateness of the collected eye-tracking data for further research.

Table 4
The results concerning the various models and fusion schemes on SumMe and ETMD databases.

	SumMe			ETMD		
	CC	NSS	AUCs	CC	NSS	AUCs
Itti_V	0.157	1.290	0.628	0.166	1.216	0.617
Itti_AV1	0.157	1.289	0.628	0.167	1.221	0.619
Itti_AV2	0.156	1.294	0.631	0.166	1.218	0.620
Itti_AV3	0.157	1.289	0.629	0.166	1.217	0.618
PQFT_V	0.095	0.874	0.586	0.101	0.816	0.577
PQFT_AV1	0.096	0.877	0.587	0.102	0.820	0.577
PQFT_AV2	0.096	0.846	0.589	0.099	0.800	0.571
PQFT_AV3	0.072	0.667	0.564	0.102	0.823	0.576
SDSR_V	0.092	0.747	0.591	0.066	0.484	0.555
SDSR_AV1	0.093	0.751	0.593	0.067	0.490	0.557
SDSR_AV2	0.098	0.805	0.605	0.083	0.619	0.576
SDSR_AV3	0.099	0.809	0.607	0.084	0.623	0.580
Deep_V	**0.194**	**1.595**	0.653	**0.254**	**1.880**	0.703
Deep_AV1	**0.194**	1.592	**0.654**	0.253	1.868	**0.704**
Deep_AV2	0.175	1.482	0.653	0.218	1.627	0.694
Deep_AV3	0.178	1.482	0.656	0.223	1.653	0.696
Min et al. (SR)	0.080	0.650	0.605	0.117	0.857	0.634

4.4.6. Analysis and discussion

We aim to analyze the performance of the various models across datasets and assess in which cases and under what circumstances the inclusion of audio indeed improves attention modeling. Regarding all databases, results on complex stimuli indicate that the audiovisual saliency model can improve eye fixation prediction results compared to the visual-only model. In some cases the improvement is small, e.g., for audiovisual PQFT on SumMe and ETMD, but in some other cases it is significant. Figs. 11, 12 present two examples from two different databases, Coutrot1 and ETMD, where the first three rows depict uniformly sampled frames from the whole video with overlayed eye-tracking data, the corresponding visual-only saliency maps and the corresponding audiovisual saliency maps. The fourth row depicts the audio waveform, while the fifth one presents the auditory saliency curve. Finally in the sixth row the visual-only saliency curve (denoted with red color) and the audiovisual saliency curve (denoted with green color) as yielded by the Deep_V and Deep_AV1 models, in terms of NSS metric are depicted. These two figures can offer an insight on how saliency evolves over time and how auditory saliency contributes to the total audiovisual saliency in our modeling. How and when auditory saliency affects visual attention has also been studied in [98].

The figures indicate that auditory saliency can reinforce the total saliency during actual audiovisual events. For example, in Fig. 11, when the related audio event begins, a boost in performance is observed, and the audiovisual saliency seems to model human attention better than visual-only saliency, an indication also reflected on eye-tracking data. Before or after the audio event, auditory saliency does not reinforce visual saliency, but at the same time, it does not degrade performance,

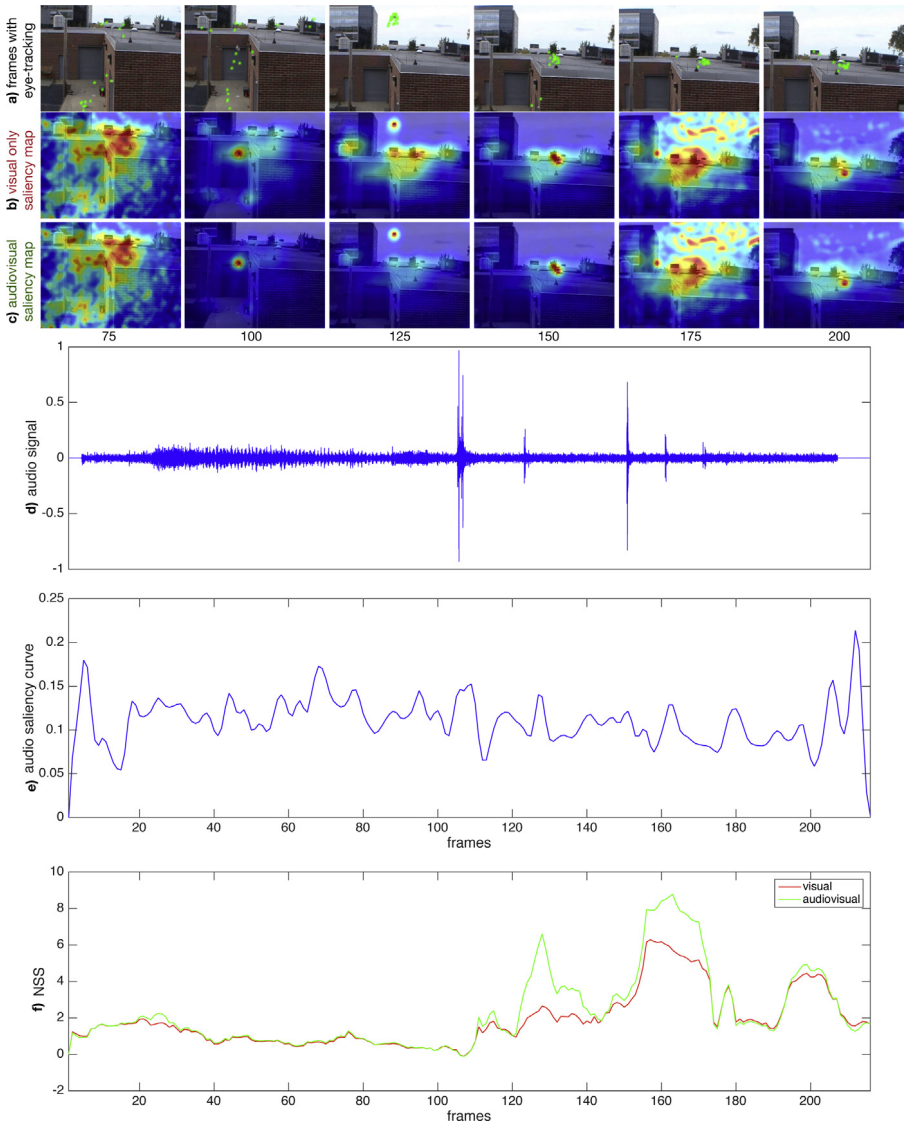

Fig. 11. An example stimulus from AVAD database. (a) Original frames with overlaid eye-tracking data are depicted along with (b) visual and (c) audiovisual saliency maps. In (d) the audio waveform and in (e) the auditory saliency curve are presented, while (f) depicts the visual-only (in red color) and the audiovisual (in green color) saliency curve in relation to NSS metric. In the beginning of the video there is noise audio, not related to the visual content. The actual audio event appears approximately in the middle of the video. Till the related audio event appears, audiovisual and visual saliencies are almost equal as seen in NSS curve (f), but when it appears, the performance of audiovisual saliency surpasses significantly the visual one. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

which is the desired behavior of the developed model. In such cases, audiovisual performance is almost equal to visual-only one. In Fig. 12 where the stimulus is more complex and the deep model has already achieved a high performance when the related audio event begins (horse galloping), still there is a small improvement depicted by the green line (audiovisual) versus the red line (visual-only). This improvement vanishes before and after the audio event, and in some cases, visual-only saliency performs slightly better than audiovisual one.

Regarding the several models, the integration of audio in Itti et al. [51] model has yielded better performance compared to the visual-only case for all databases. We performed some indicative ANOVA tests to confirm the statistical significance of our results. In Coutrot1 database, ANOVA test between Itti_V and Itti_AV1 yields an F-statistic of $F = 8.267$ ($p<0.005$) for CC, while for Coutrot2 it yields $F = 3810$ ($p<0.001$) for NSS. For all databases, the developed deep models yield the best absolute results, in most cases with audiovisual fusion increasing the performance compared to visual-only saliency. In many cases, Deep_V and Deep_AV1 are very close, thus we perform ANOVA tests again to assess the statistical significance of these results. For Coutrot1 database the ANOVA test yields $F = 9.5113$ ($p<0.005$), 52.822 ($p<0.001$), $F = 9.890$ ($p<0.001$) for CC, NSS, AUCs and for Coutrot2 $F = 32.642$ ($p<0.001$), $F = 43.941$ ($p<0.001$), $F = 5.518$ ($p<0.05$) respectively. In ETMD database the corresponding results for Deep_V and Deep_AV1 are $F = 85.284$ ($p<0.001$), $F = 65.350$ ($p<0.001$) and $F = 4.726$ ($p<0.05$).

Regarding the comparison with the employed state-of-the-art audiovisual model, Min et al. method [11] performs second best in AVAD database [11], but its performance in terms of CC and NSS is lowered when applied on more complex stimuli of longer duration, like movies, that might contain very few actual audiovisual events. Regarding AUCs results though, Min et al. performance is better than Itti_AV1, and comparable to Deep models, especially in DIEM and ETMD databases.

5. Conclusion

We have developed a computational audiovisual saliency model based on behaviorally-inspired fusion schemes between well-known individual saliency models and aspire to validate its plausibility via human behavioral experiments and eye-tracking data. We propose three

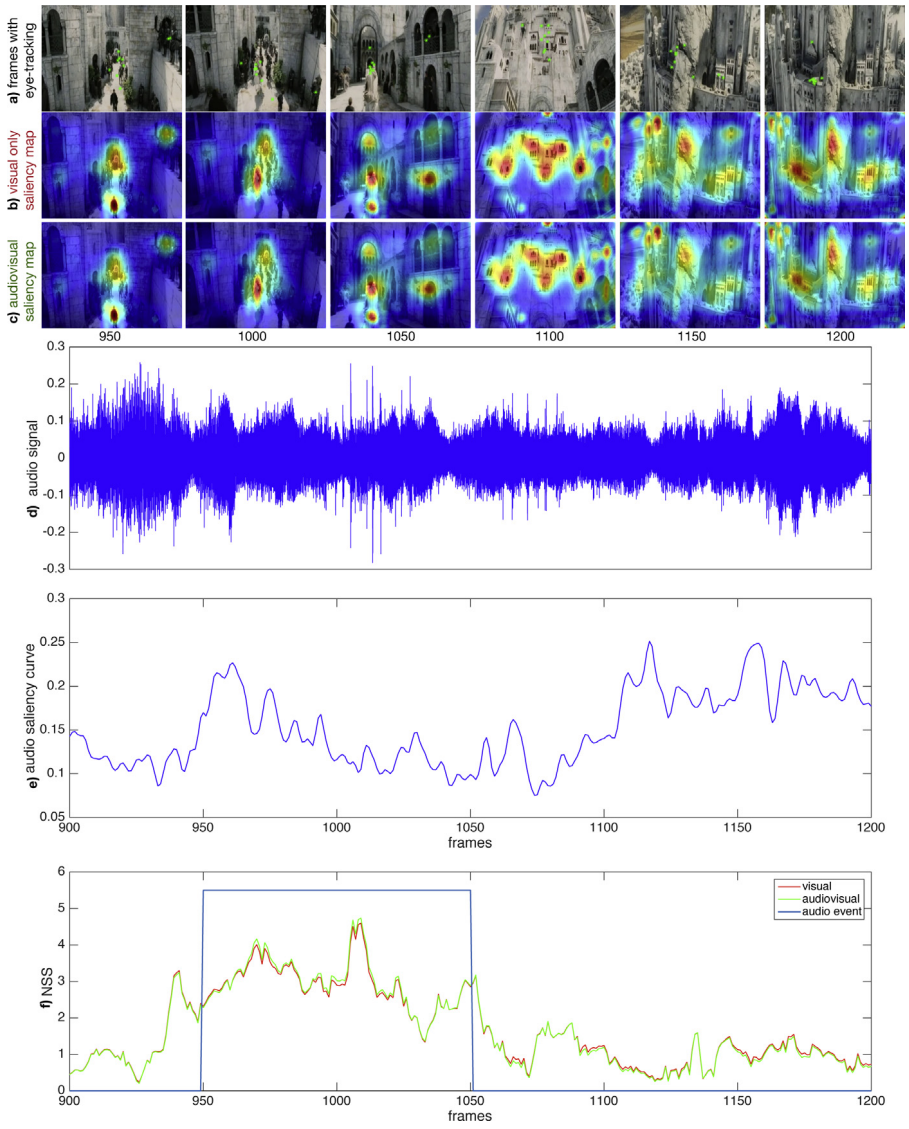

Fig. 12. An example stimulus from ETMD database. (a) Original frames with overlaid eye-tracking data are depicted along with (b) visual and (c) audiovisual saliency maps. In (d) the audio waveform and in (e) the auditory saliency curve are presented, while (f) depicts the visual-only (in red color) and the audiovisual (in green color) saliency curve in relation to NSS metric. Also the blue rectangle indicates the duration of an audiovisual event (horse galopping). Although the differences between audiovisual and visual saliencies here are small, we can still notice that during the audio event, NSS metric for audiovisual saliency is slightly better, while before and after the event, they are almost equal or alternating between better and worse. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

fusion schemes and subsequently evaluate them. Our first validation effort concerns the "pip and pop" and "sine vs. square" effects, where our model exhibits a similar behavior to the experimental results compared to visual-only models. Regarding the second evaluation strategy, with human audiovisual eye-tracking data, we assess the performance of the several fusion schemes and saliency models on six different databases of variable complexity, DIEM, AVAD, Coutrot1, Coutrot2, SumMe, and ETMD. For SumMe and ETMD, we have collected audiovisual eye-tracking data which we are going to publicly release, which is another contribution of this work. Results for both evaluation strategies and across multiple datasets are promising and indicate the superiority of audiovisual saliency versus visual-only one, even in complex stimuli.

Acknowledgments

This work was cofinanced by the European Regional Development Fund of the EU and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call 'Research - Create - Innovate' (T1EDK-01248, "i-Walk").

The authors wish to thank all the members of the NTUA CVSP Lab who participated in the audiovisual eye-tracking data collection. Special thanks to Efthymios Tsilionis for sharing his code and his advice during eye-tracking database collection.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.image.2019.05.001.

References

[1] M.A. Meredith, B.E. Stein, Interactions among converging sensory inputs in the superior colliculus, Science 221 (4608) (1983) 389–391.
[2] M.A. Meredith, B.E. Stein, Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration, J. Neurophysiol. 56 (3) (1986) 640–662.
[3] A. Vatakis, C. Spence, Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli, Percept. Psychophys. 69 (5) (2007) 744–756.
[4] P. Maragos, A. Gros, A. Katsamanis, G. Papandreou, Cross-modal integration for performance improving in multimedia: A review, in: Multimodal Processing and Interaction: Audio, Video, Text, Springer-Verlag, 2008, pp. 1–46.

[5] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature 264 (1976) 746–748.

[6] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention, Vis. Res. 42 (1) (2002) 107–123.

[7] E. Van der Burg, C.N.L. Olivers, A.W. Bronkhorst, J. Theeuwes, Pip and pop: Nonspatial auditory signals improve spatial visual search, J. Exp. Psychol. Hum. Percept. Perform. 34 (2008) 1053–1065.

[8] C. Koch, S. Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, Hum. Neurobiol. 4 (1985) 219–227.

[9] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[10] C. Kayser, C.I. Petkov, M. Lippert, N.K. Logothetis, Mechanisms for allocating auditory attention: An auditory saliency map, Curr. Biol. 15 (21) (2005) 1943–1947.

[11] X. Min, G. Zhai, K. Gu, X. Yang, Fixation prediction through multimodal analysis, ACM Trans. Multimed. Comput. Commun. Appl. 13 (1) (2017).

[12] A. Coutrot, N. Guyader, An audiovisual attention model for natural conversation scenes, in: Proc. IEEE Int. Conf. on Image Processing, 2014, pp. 1100–1104.

[13] G. Potamianos, C. Neti, G. Gravier, A. Garg, A.W. Senior, Recent advances in the automatic recognition of audiovisual speech, Proc. IEEE 91 (9) (2003) 1306–1326.

[14] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, Y. Avrithis, Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention, IEEE Trans. Multimed. 15 (7) (2013) 1553–1568.

[15] G. Schillaci, S. Bodiroža, V.V. Hafner, Evaluating the effect of saliency detection and attention manipulation in human-robot interaction, Int. J. Soc. Robot. 5 (1) (2013) 139–152.

[16] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, P. Maragos, Multimodal human action recognition in assistive human-robot interaction, in: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing, 2016, pp. 2702–2706.

[17] A. Tsiami, P. Koutras, N. Efthymiou, P.P. Filntisis, G. Potamianos, P. Maragos, Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots, in: Int. Conf. on Robotics and Automation, 2018, pp. 4585–4592.

[18] A. Tsiami, A. Katsamanis, P. Maragos, A. Vatakis, Towards a behaviorally-validated computational audiovisual saliency model, in: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing, 2016, pp. 2847–2851.

[19] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, R. Pfeifer, Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub, in: Int. Conf. on Robotics and Automation, 2008, pp. 962–967.

[20] B. Schauerte, B. Kühn, K. Kroschel, R. Stiefelhagen, Multimodal saliency-based attention for object-based scene analysis, in: IEEE Int. Conf. on Intelligent Robots and Systems, IEEE, 2011, pp. 1173–1179.

[21] S. Ramenahalli, D.R. Mendat, S. Dura-Bernal, E. Culurciello, E. Nieburt, A. Andreou, Audio-visual saliency map: Overview, basic models and hardware implementation, in: Proc. Information Sciences and Systems, CISS, 2013, pp. 1–6.

[22] R. Ratajczak, D. Pellerin, Q. Labourey, C. Garbay, A fast audiovisual attention model for human detection and localization on a companion robot, in: Int. Conf. on Applications and Systems of Visual Paradigms, 2016.

[23] N. Sidaty, M.-C. Larabi, A. Saadane, Toward an audiovisual attention model for multimodal video content, Neurocomputing (2017) 94–111.

[24] Y. Chen, T.V. Nguyen, M.S. Kankanhalli, J. Yuan, S. Yan, M. Wang, Audio matters in visual attention., IEEE Trans. Circuits Syst. Video Technol. 24 (11) (2014) 1992–2003.

[25] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, Y. Avrithis, Video event detection and summarization using audio, visual and text saliency, in: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing, 2009, pp. 3553–3556.

[26] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, A. Potamianos, Predicting audio-visual salient events based on visual, audio and text modalities for movie summarization, in: Proc. IEEE Int. Conf. on Image Processing, 2015, pp. 4361–4365.

[27] A. Coutrot, N. Guyader, How saliency, faces, and sound influence gaze in dynamic social scenes, J. Vis. 14 (8) (2014) 1–17.

[28] A. Coutrot, N. Guyader, Multimodal saliency models for videos, in: M. Mancas, V.P. Ferrera, N. Riche, J.G. Taylor (Eds.), From Human Attention to Computational Attention: A Multidisciplinary Approach, Springer, New York, 2016, pp. 291–304.

[29] G. Song, Effect of Sound in Videos on Gaze: Contribution to Audio-Visual Saliency Modeling (Ph.D. thesis), Universite de Grenoble, 2013.

[30] X. Min, G. Zhai, Z. Gao, C. Hu, X. Yang, Sound influences visual attention discriminately in videos, in: 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), IEEE, 2014, pp. 153–158.

[31] X. Min, G. Zhai, C. Hu, K. Gu, Fixation prediction through multimodal analysis, in: Proc. IEEE Int. Conf. on Visual Communications and Image Processing.

[32] S. Morein-Zamir, S. Soto-Faraco, A. Kingstone, Auditory capture of vision: Examining temporal ventriloquism, Cogn. Brain Res. 17 (1) (2003) 154–163.

[33] W. Fujisaki, S. Shimojo, M. Kashino, S. Nishida, Recalibration of audiovisual simultaneity, Nature Neurosci. 7 (7) (2004) 773–778.

[34] E. Van der Burg, D. Alais, J. Cass, Rapid recalibration to audiovisual asynchrony, J. Neurosci. 33 (37) (2013) 14633–14637.

[35] E. Van der Burg, E. Orchard-Mills, D. Alais, Rapid temporal recalibration is unique to audiovisual stimuli, Exp. Brain Res. 233 (1) (2015) 53–59.

[36] M. Keetels, J. Vroomen, Sound affects the speed of visual processing, J. Exp. Psychol. Hum. Percept. Perform. 37 (3) (2011) 699–708.

[37] S. Gleiss, C. Kayser, Eccentricity dependent auditory enhancement of visual stimulus detection but not discrimination, Front. Integr. Neurosci. 7 (52) (2013) 1–8.

[38] Q. Li, H. Yang, F. Sun, J. Wu, Spatiotemporal relationships among audiovisual stimuli modulate auditory facilitation of visual target discrimination, Percept. Abstr. 44 (3) (2015) 232–242.

[39] D. Burr, M.S. Banks, M.C. Morrone, Auditory dominance over vision in the perception of interval duration, Exp. Brain Res. 198 (1) (2009) 49–57.

[40] L. Chen, J. Vroomen, Intersensory binding across space and time: A tutorial review, Atten. Percept. Psychophys. 75 (5) (2013) 790–811.

[41] M.O. Ernst, A Bayesian view on multimodal cue integration, Hum. Body Percept. Inside Out (2006) 105–131.

[42] D. Alais, D. Burr, The ventriloquist effect results from near-optimal bimodal integration, Curr. Biol. 14 (3) (2004) 257–262.

[43] H. Colonius, A. Diederich, The optimal time window of visual-auditory integration: A reaction time analysis, Front. Integr. Neurosci. 4 (11) (2010) 1–8.

[44] V. van Wassenhove, K.W. Grant, D. Poeppel, Temporal window of integration in auditory-visual speech perception, Neuropsychologia 45 (3) (2007) 598–607.

[45] A. Borji, L. Itti, State-of-the-art in visual attention modeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 185–207.

[46] A.M. Treisman, G. Gelade, A feature-integration theory of attention, Cogn. Psychol. 12 (1) (1980) 97–136.

[47] R. Milanese, Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation (Ph.D. thesis), University of Geneva, 1993.

[48] S. Baluja, D. Pomerleau, Using a saliency map for active spatial selective attention: Implementation & initial results, in: Advances in Neural Information Processing Systems, 1994, pp. 451–458.

[49] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y. Lai, N. Davis, F. Nuflo, Modeling visual attention via selective tuning, Artificial Intelligence 78 (1–2) (1995) 507–545.

[50] E. Niebur, C. Koch, Control of selective visual attention: Modeling the where pathway, in: Advances in Neural Information Processing Systems, 1995, pp. 802–808.

[51] L. Itti, N. Dhavale, F. Pighin, Realistic avatar eye and head animation using a neurobiological model of visual attention, in: Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI, vol. 5200, International Society for Optics and Photonics, 2003, pp. 64–79.

[52] D. Walther, C. Koch, Modeling attention to salient proto-objects, J. Neural Netw. 19 (9) (2006) 1395–1407.

[53] S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search, in: Lecture Notes in Computer Science, vol. 3899, Springer, 2006.

[54] O.L. Meur, P.L. Callet, D. Barba, D. Thoreau, A coherent computational approach to model bottom-up visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 28 (5) (2006) 802–817.

[55] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, in: Advances in Neural Information Processing Systems, 2007, pp. 545–552.

[56] O.L. Meur, P.L. Callet, D. Barba, Predicting visual fixations on video based on low-level visual features, Vis. Res. 47 (19) (2007) 2483–2498.

[57] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué, Modelling spatio-temporal saliency to predict gaze direction for short videos, Int. J. Comput. Vis. 82 (3) (2009) 231–243.

[58] K. Rapantzikos, Y. Avrithis, S. Kollias, Spatiotemporal features for action recognition and salient event detection, in: J.G. Taylor, V. Cutsuridis (Eds.), Cogn. Comput. 3 (1) (2011) 167–184, special issue on Saliency, attention, visual search and picture scanning.

[59] A. Garcia-Diaz, X.R. Fernandez-Vidal, X.M. Pardo, R. Dosil, Saliency from hierarchical adaptation through decorrelation and variance normalization., Image Vis. Comput. 30 (1) (2012) 51–64.

[60] A. Torralba, Modeling global scene factors in attention, J. Opt. Soc. Amer. A 20 (2003) 1407–1418.

[61] L. Itti, P. Baldi, Bayesian surprise attracts human attention, in: Proc. Advances in Neural Information Processing Systems, 2005.

[62] I. Gkioulekas, G. Evangelopoulos, P. Maragos, Spatial Bayesian surprise for image saliency and quality assessment, in: Proc. IEEE Int. Conf. on Image Processing, 2010.

[63] A. Oliva, A. Torralba, M.S. Castelhano, J.M. Henderson, Top-down control of visual attention in object detection, in: Proc. IEEE Int. Conf. on Image Processing, 2003.

[64] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell, Sun: A Bayesian framework for saliency using natural statistics, J. Vis. 8 (7) (2008) 1–20.

[65] L. Zhang, M.H. Tong, G. W., Sunday: Saliency using natural statistics for dynamic analysis of scenes, in: Proc. Cognitive Science Society Conference, 2009, pp. 2944–2949.
[66] T. Kadir, M. Brady, Saliency, scale and image description, Int. J. Comput. Vis. 45 (2) (2001) 83–105.
[67] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Advances in Neural Information Processing Systems, 2006, pp. 155–162.
[68] X. Hou, L. Zhang, Dynamic visual attention: Searching for coding length increments, in: Advances in Neural Information Processing Systems, 2009, pp. 681–688.
[69] D. Gao, N. Vasconcelos, Discriminant saliency for visual recognition from cluttered scenes, in: Advances in Neural Information Processing Systems, 2004, pp. 481–488.
[70] D. Gao, S. Han, N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 989–1005.
[71] H.J. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vis. 9 (12) (2009) 1–27.
[72] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, T. Dutoit, Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis, Signal Process., Image Commun. 28 (6) (2013) 642–658.
[73] N. Bruce, J. Tsotsos, Spatiotemporal saliency: Towards a hierarchical representation of visual saliency, in: Int'l Workshop on Attention and Performance in Comp. Vis., 2008.
[74] V. Mahadevan, N. Vasconcelos, Spatiotemporal saliency in dynamic scenes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 171–177.
[75] X. Hou, L. Zhang, Saliency detection: A spectral residual approach, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[76] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2009, pp. 1597–1604.
[77] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Trans. Image Process. 19 (1) (2010) 185–198.
[78] X. Hou, J. Harel, C. Koch, Image signature: Highlighting sparse salient regions, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 194–201.
[79] B. Schauerte, R. Stiefelhagen, Quaternion-based spectral saliency detection for eye fixation prediction, in: Proc. European Conf. on Computer Vision, 2012, pp. 116–129.
[80] C. Guo, Q. Ma, L. Zhang, Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[81] M. Mancas, N. Riche, J. Leroy, B. Gosselin, Abnormal motion selection in crowds using bottom-up saliency, in: Proc. IEEE Int. Conf. on Image Processing, 2011, pp. 175–178.
[82] T.V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, S. Yan, Static saliency vs. dynamic saliency: A comparative study, in: Proc. ACM Int. Conf. on Multimedia, 2013, pp. 987–996.
[83] P. Koutras, P. Maragos, A perceptually based spatio-temporal computational framework for visual saliency estimation, Signal Process., Image Commun. 38 (2015) 15–31.
[84] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, A coherent computational approach to model bottom-up visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 28 (5) (2006) 802–817.
[85] E. Vig, M. Dorr, D. Cox, Large-scale optimization of hierarchical features for saliency prediction in natural images, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2014, pp. 2798–2805.
[86] M. Kümmerer, L. Theis, M. Bethge, Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet, in: Int. Conf. on Learning Representations Workshop, 2015.
[87] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, N.E. O'Connor, Shallow and deep convolutional networks for saliency prediction, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2016, pp. 598–606.
[88] X. Huang, C. Shen, X. Boix, Q. Zhao, Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks, in: Proc. IEEE Int. Conf. on Computer Vision, 2015, pp. 262–270.
[89] N. Liu, J. Han, D. Zhang, S. Wen, T. Liu, Predicting eye fixations using convolutional neural networks, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2015, pp. 362–370.
[90] S. Jetley, N. Murray, E. Vig, End-to-end saliency mapping via probability distribution prediction, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2016, pp. 5753–5761.
[91] J. Pan, C. Canton, K. McGuinness, N.E. O'Connor, J. Torres, E. Sayrol, X.a. Giro-i Nieto, Salgan: Visual saliency prediction with generative adversarial networks, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition Workshop, 2017.
[92] C. Bak, A. Kocak, E. Erdem, A. Erdem, Spatio-temporal saliency networks for dynamic saliency prediction, IEEE Trans. Multimed. (2017).
[93] G. Leifman, D. Rudoy, T. Swedish, E. Bayro-Corrochano, R. Raskar, Learning gaze transitions from depth to improve video saliency estimation, in: Proc. IEEE Int. Conf. on Computer Vision, 2017.
[94] E.M. Kaya, M. Elhilali, A temporal saliency map for modeling auditory attention, in: Proc. Information Sciences and Systems, CISS, 2012, pp. 1–6.
[95] O. Kalinli, S.S. Narayanan, A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech, in: Proc. Int. Conf. on Speech Communication and Technology, 2007, pp. 1941–1944.
[96] L. Itti, C. Koch, Computational modelling of visual attention, Nat. Rev. Neurosci. 2 (3) (2001) 194–203.
[97] B. Schauerte, R. Stiefelhagen, "wow!" Bayesian surprise for salient acoustic event detection, in: Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing, 2013, pp. 6402–6406.
[98] A. Coutrot, N. Guyader, G. Ionescu, A. Caplier, Video viewing: Do auditory salient events capture visual attention?, Ann. Telecommun. 69 (1) (2014) 89–97.
[99] C. Bordier, F. Puja, E. Macaluso, Sensory processing during viewing of cinematographic material: Computational modeling and functional neuroimaging, Neuroimage 67 (2013) 213–226.
[100] J.W. Gebhard, G.H. Mowbray, On discriminating the rate of visual flicker and auditory flutter, Am. J. Psychol. 72 (4) (1959) 521–529.
[101] S. Shimojo, L. Shams, Sensory modalities are not separate modalities: Plasticity and interactions, Curr. Opin. Neurobiol. 11 (4) (2001) 505–509.
[102] Y. Wada, N. Kitagawa, K. Noguchi, Audio–visual integration in temporal perception, Int. J. Psychophysiol. 50 (1) (2003) 117–124.
[103] L. Shams, Y. Kamitani, S. Shimojo, What you see is what you hear, Nature 408 (2000) 788.
[104] R.B. Welch, L.D. DutionHurt, D.H. Warren, Contributions of audition and vision to temporal rate perception, Percept. Psychophys. 39 (4) (1986) 294–300.
[105] R. Sekuler, A.B. Sekuler, R. Lau, Sound alters visual motion perception, Nature 385 (6614) (1997) 308.
[106] E. Van der Burg, J. Cass, C.N.L. Olivers, J. Theeuwes, D. Alais, Efficient visual search from synchronized auditory signals requires transient audiovisual events, PLoS One 5 (5) (2010) e10664.
[107] C.V. Parise, V. Harrar, M.O. Ernst, C. Spence, Cross-correlation between auditory and visual signals promotes multisensory integration, Multisens. Res. 26 (3) (2013) 307–316.
[108] M. Rolf, M. Hanheide, K.J. Rohlfing, Attention via synchrony: Making use of multimodal cues in social learning, IEEE Trans. Auton. Ment. Dev. 1 (1) (2009) 55–67.
[109] J. Hershey, J. Movellan, Audio-vision: Using audio-visual synchrony to locate sounds, Adv. Neural Inf. Process. Syst. 12 (2000) 813–819.
[110] A. Borji, D.N. Sihite, L. Itti, Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study, IEEE Trans. Image Process. 22 (1) (2013) 55–69.
[111] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, What do different evaluation metrics tell us about saliency models?, 2016, arXiv preprint arXiv:1604.03605.
[112] L. Zhaoping, L. Zhe, Primary visual cortex as a saliency map: A parameter-free prediction and its test by behavioral data, PLoS Comput. Biol. 11 (10) (2015) 1–39.
[113] H.J. Müller, P.M. Rabbitt, Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption, J. Exp. Psychol. Hum. Percept. Perform. 15 (2) (1989) 315–330.
[114] M. Gygli, H. Grabner, H. Riemenschneider, L. Van Gool, Creating summaries from user videos, in: Proc. European Conf. on Computer Vision, 2014, pp. 505–520.
[115] P.K. Mital, T.J. Smith, R. Hill, J.M. Henderson, Clustering of gaze during dynamic scene viewing is predicted by motion, Cogn. Comput. 3 (1) (2011) 5–24.
[116] P. Koutras, A. Katsamanis, P. Maragos, Predicting eyes fixations in movie videos: Visual saliency experiments on a new eye-tracking database, in: Proc. Human Computer Interaction Conf. (Eng. Psychology and Cognitive Ergonomics), 2014, pp. 183–194.
[117] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: Towards good practices for deep action recognition, in: Proc. European Conf. on Computer Vision, 2016, pp. 20–36.
[118] C. Zach, T. Pock, H. Bischof, A duality based approach for realtime tv-l 1 optical flow, in: Joint Pattern Recognition Symposium, Springer, 2007, pp. 214–223.
[119] D. Rudoy, D.B. Goldman, E. Shechtman, L. Zelnik-Manor, Learning video saliency from human gaze using candidate selection, in: Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2013, pp. 1147–1154.
[120] W. Wang, J. Shen, F. Guo, M.-M. Cheng, A. Borji, Revisiting video saliency: A large-scale benchmark and a new model, 2018, arXiv preprint arXiv:1801.07424.