Audio-Visual Temporal Saliency Modeling Validated by fMRI Data

Petros Koutras, Georgia Panagiotaropoulou, Antigoni Tsiami, and Petros Maragos
School of E.C.E., National Technical University of Athens, Greece
{pkoutras, antsiami, maragos}@cs.ntua.gr, gio.panagiotaropoulou@gmail.com

Abstract

In this work we propose an audio-visual model for predicting temporal saliency in videos, that we validate and evaluate in an alternative way by employing fMRI data. We intend to bridge the gap between the large improvements achieved during the last years in computational modeling, especially in deep learning, and the neurobiological and behavioral research regarding human vision. The proposed audio-visual model incorporates both state-of-the-art deep architectures for visual saliency, which were trained on eye-tracking data, and behavioral findings concerning audio-visual integration in multimedia stimuli. A new fMRI database has been collected for evaluation purposes, that includes various videos and subjects. This dataset may prove useful not only for saliency but for other computer vision problems as well. The evaluation of our model using the new fMRI database under a mixed-effect analysis shows that the proposed saliency model has strong correlation with both the visual and audio brain areas, that confirms its effectiveness and appropriateness in predicting audio-visual saliency for dynamic stimuli.

1. Introduction

Nowadays, the breakthrough in the area of deep learning is revolutionizing many fields in the area of computer vision. The extensive usage of Convolutional Neural Networks (CNNs) has boosted the performance throughout the majority of tasks in computer vision, such as object detection or semantic segmentation [60, 34, 23]. One of the major downsides of deep network approaches is their need for large-scale training datasets. In image domain, many approaches employ pre-trained network architectures trained on ImageNet [32] for object classification, or SALICON [31] for static saliency estimation. However, the progress of CNN architectures, design, and representation learning in the video domain is much slower, and the performance of deep learning methods remains comparable with non-deep ones. The main difficulties arise both from the lack of large-scale video datasets, and the way of integrating temporal information in a deep architecture, i.e., the best method of temporal aggregation in video (recurrent vs convolutional).

Among the video domain related problems, dynamic saliency estimation is most closely related to brain neural responses, since various stages of biological vision systems involve spatio-temporal processing, and nature has a tendency to represent information in optimal ways. Visual saliency is a bottom-up process and is based on the sensory cues of a stimulus that make certain image or video regions more conspicuous. During the last years, various computational approaches have been developed for visual saliency estimation in the spatial domain. Several among them have already incorporated advances from the deep learning research. In parallel, spatio-temporal models for saliency estimation in video stimuli have also appeared, but their performance remains slightly better or only competitive compared to the best static saliency approaches [6, 68].

Generally, the modeling of video saliency can be approached by two different representations: The first consists of spatio-temporal saliency maps employed for the task of dynamic fixation prediction in videos. In the second representation, the produced spatio-temporal maps are mapped to a 1D map yielding time-varying saliency curves. These curves can be used in a video summarization task, since they can be viewed as an indicator function that describes the interestingness of each frame in a video sequence [14, 13]. In our work, we take advantage of the existing approaches for eye-fixation prediction, especially the deep models trained on big eye-tracking databases, and propose a method that transforms the produced saliency maps to 1D temporal saliency curves.

Multisensory interaction and integration in the human brain manifest themselves in multiple ways and in multiple contexts [66, 42]. There is considerable evidence that human attention is influenced by multimodal and specially audio-visual information [44, 45]. In addition, video data are in general multimodal, containing visual, audio and semantic streams, and of particular interest is the estimation of a multimodal temporal saliency curve that models human attention during a video viewing. The works of [14, 36] proposed a multimodal framewise saliency model

based on visual, audio and text cues which has been integrated in a multimodal system for movie summarization. Moreover, [63] proposed a behaviorally validated 2D audio-visual saliency model that is able to explain behavioral experiments in video stimuli. In our study we propose fusion strategies in order to integrate the audio information in a temporal audio-visual model, where visual saliency is modulated by audio saliency. We are interested not only in developing such a model, but also in validating its plausibility with human data.

Towards this goal, in parallel with research in computational modeling and machine learning, brain imaging techniques such as functional Magnetic Resonance Imaging (fMRI), can serve as a noninvasive tool to monitor neural activity during external stimulation, thus illuminating the structural and functional architecture of the human brain. Recently, there has been a shift towards more complex and naturalistic stimuli, such as real-life images, video and audio excerpts. The attempt to study such real-life stimuli aims at understanding their representation in the human brain, and ultimately at linking low-level features with the high-level semantic information they convey, in order to propose and improve computational models for many computer vision tasks [39]. The presentation of videos with simultaneous acquisition of fMRI data provides a semi-natural setup to infer the complex mechanisms employed by the human brain to represent and comprehend such stimuli, while at the same time posing a challenge to develop efficient as well as cognitively plausible computational designs to model the underlying neural processes.

In this work we try to bridge the gap between the huge progress in the computational approaches for computer vision, and the neurobiological and psychophysical evidences about human vision obtained by analyzing fMRI data. Our goal is to build an audio-visual temporal saliency model and validate its plausibility through fMRI data. This validation aims at confirming that our model indeed captures the behavior of human audio-visual attention when exposed to audio-visual stimuli. The validation process essentially corresponds to investigating whether fMRI data exhibit activation in the areas that care expected to get activated when humans are exposed to specific stimuli.

One interesting question is to what extent human individuals have the same perception of identical stimuli presented to them and whether the neural representations they create are fundamentally different or share the same structure. If the latter proves to be the case, brain imaging data could be used to augment computational models and further the deep learning representation of multimedia in accordance to human perception. For this purpose we have collected a large amount of fMRI data using video stimuli viewed by multiple persons. This dataset could be useful for many computer vision problems related to the video domain such as dynamic saliency, object and action recognition or movie summarization. In addition, due to the fact that the proposed dataset contains both multiple videos and subjects, with a proper statistical analysis we could generalize our observations from the specific samples to the entire underlying population.

The contributions of the paper can be summarized as follows:

- First, we propose an audio-visual temporal saliency model, in the form of a temporal saliency curve instead of the most commonly used saliency map, for predicting saliency in videos. This approach is based on the modification of state-of-the-art methods for visual saliency, and additionally it incorporates audio information using different fusion schemes (Section 3).

- Second, a new fMRI database has been collected, that contains both multiple video stimuli and multiple subjects. This dataset can be useful for evaluating and improving many computational methods for video-based computer vision tasks, and also for understanding how these methods are related to processes in human brain (Section 4).

- Third, the proposed audio-visual temporal saliency model is evaluated using the collected fMRI dataset, and the results indicate that this model has strong correlation with both the visual and audio brain areas. In addition, unlike previous studies [4, 51] where only one movie was employed, we apply a mixed effect analysis, which gives stronger confidence and allows the generalization of our results to the general population. (Sections 5 and 6).

2. Related Work

2.1. Visual Saliency Models

Visual saliency constitutes one of the most important problems in both cognitive and computer vision, and many methods have been developed for saliency prediction, especially for still images, i.e. spatial-only methods [62, 6, 5]. Regarding spatio-temporal saliency, less work has been done compared to spatial-only, and in most cases the existing spatio-temporal models are an extension of spatial ones, by incorporating additional dynamic visual features. For example, in [28, 27, 21] differences between the spatial orientation maps are employed as temporal features for saliency detection in videos, while [7, 71, 24, 41] take advantage of features statistics computed on dynamic stimuli. In [35] a perceptually based spatio-temporal computational framework for visual saliency estimation is presented, based on quadrature Gabor filters in three dimensions. In [57] the authors extend their self-resemblance method by employing

3D local steering kernels for action and saliency detection in videos. In another class of approaches, saliency is estimated in the frequency domain by employing the quaternion Fourier transform for color, intensity and motion features [19, 20].

During the very last years, a large amount of works approach the problem of visual saliency by employing deep neural networks. Some approaches are based on the adaptation of pretrained CNN models for visual recognition tasks [37], while in [50] both shallow and deep CNN are trained end-to-end for saliency prediction. In [26], multiscale information is employed for training CNN networks by optimizing common saliency evaluation metrics while the work of [30] showed that losses based on probability distance measures may be more suitable for saliency rather than standard loss functions for regression. In [3] the authors proposed a two-stream CNN network based on RGB images and optical flow maps for dynamic saliency prediction. In [38], gaze transitions are learned from RGB, optical flow and depth information in order to improve saliency estimation in videos.

2.2. Audio-Visual Integration in Saliency

However, our daily experience as well as systematic behavioral experiments indicate the strong audio-visual interactions that draw our audio-visual attention. Well-known examples of strong audio-visual interactions are the McGurk effect [43], or the bouncing ball illusion [59]. Several attempts to model audio-visual attention exist in the literature, but most of them are application-specific or use spatial audio in order to fuse it with visual information, e.g., in robotic applications. A computational audio-visual saliency model that predicts attention in an audiovisual scene, i.e., where the eye are be fixated, has for the first time been presented in [54] and has been developed to guide a humanoid robot. In this model, estimation of visual saliency is based on the Itti et al. approach [29], while for audio, only the spatial properties of the sources are integrated. Similarly, in [52], the auditory saliency map is also estimated via source localization and then fused with visual saliency via a product operation. The model proposed in [55] is also based on source localization, but also on Bayesian surprise for auditory saliency map generation, and on a phase-based approach for visual saliency. For a slightly different application, the audiovisual model introduced in [15, 14] aims primarily to summarize movies or videos. This work has been further improved in [36]. Both models aim at predicting when, and not where, attention would be drawn in a dynamic scene. All the above mentioned models are primarily application-oriented and despite having possibly been inspired by cognitive science, no effort has been made to validate their behavior in comparison to behavioral findings.

Coutrot and Guyader [10, 11] as well as Song [61] have tried to more directly validate their models with humans with their findings indicating that, in movies, eye gaze is attracted by talking faces and music players. The model presented in [63] focuses on predicting audio-visual saliency in videos, by appropriately combining existing auditory and visual saliency models in order to form an audio-visual saliency model that is also behaviorally validated. Subsequently, the audio-visual model is compared against findings from behavioral experiments.

2.3. FMRI analysis on Multimedia Data

The most reliable validation strategy for all computational saliency methods is through comparison with actual human data. Several contributions have so far been made towards linking computational frameworks to brain activation data. Such efforts aim at establishing new methods of combining and interpreting the two types of data [8, 46, 12], at assessing the biological plausibility of widespread perceptual models [4, 72] or at augmenting the latter by integrating high-level information encoded inside the human brain [25, 40]. Another study proposes that whenever different individuals are exposed to the same audiovisual stimulus, the internal brain representations they form should be similar, since they encode information (features) of the stimulus itself. Thus, brain regions involved in audiovisual processing should have similar time responses across individuals, in contrast to others [22].

3. Audio-Visual Model for Temporal Saliency

As briefly described in the introduction, our goal is to create an audio-visual saliency model, able to predict temporal saliency. For this purpose, we employ existing saliency models that have been proposed for the fixation prediction problem, and have been trained with large-scale eye-tracking databases. In our approach we modify and extend these methods in order to deal with the problem of audio-visual temporal saliency prediction in videos, without any additional training. We essentially intend to transfer the knowledge from the eye-tracking databases to a closely related problem. The main important parts and parameters of our model that have to be defined and designed are the following: 1) the decision of where the static and the temporal components of the visual saliency model will be fused (using 2D saliency maps or 1D saliency curves), 2) the type of the fusion scheme and, 3) how audio information will be integrated in the audiovisual model. The employed approach is depicted in Fig. 1, and is analyzed further in the following sections.

3.1. Visual Model for Temporal Saliency

For the visual saliency modeling, we follow a hybrid approach that incorporates a state-of-the-art CNN network for static saliency, and an optical flow estimation as the temporal saliency component. We did not employ a fully deep-

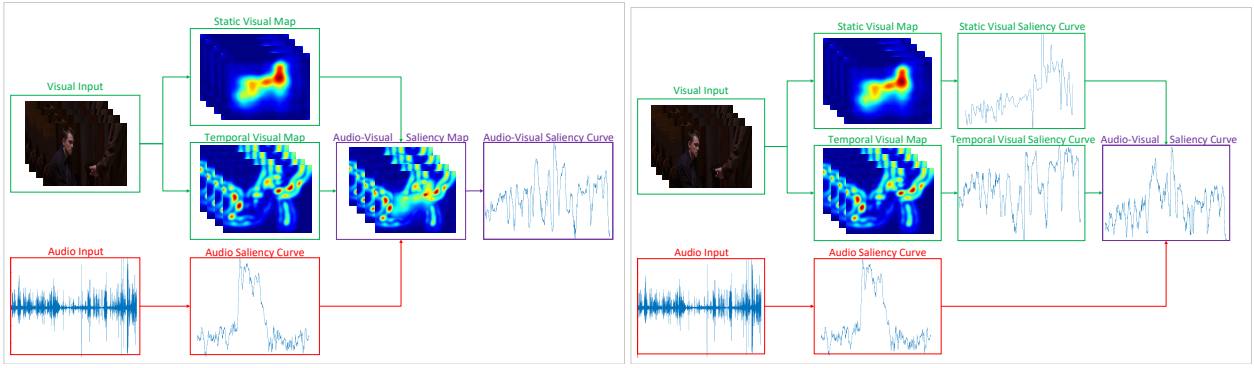


Figure 1: Overview of the audio-visual temporal saliency model: a) fusion in the level of saliency maps, b) fusion in the level of saliency curves.

based spatio-temporal network, such as [68], because we needed to maintain the two components separately, in order to investigate the different fusion approaches and incorporate the audio information as well.

3.1.1 Static Component

For the static component we used the publicly available deep model from [50]. The architecture of this network is identical to a VGG-M network. However, the authors have replaced the three fully connected layers with convolutional layers in order to make the network structure suitable for the task of saliency estimation. In addition, a deconvolution layer is employed as the final layer in order to resize the output to the image size. The network was trained on 9000 images from SALICON dataset using an Euclidean loss function. The output of the network constitutes the static saliency map M_S, with values in $[0, 1]$.

3.1.2 Temporal Component

For temporal saliency, we extract warped optical flow maps using the implementation of [67], which is based on the TVL1 optical flow algorithm [70]. Then, a temporal moving averaging filter over ten successive frames is applied to smooth and remove the noise from optical flow estimation in x and y directions independently. Afterwards, we apply Difference-of-Gaussians (DoG) filtering to the optical flow magnitude as in [53]. Since the resulting saliency map has small values with a few noisy spikes, we use logarithm in order to suppress these sharp peaks. Finally, we normalize the temporal saliency map M_T with its maximum value across all video frames.

3.1.3 From 2D Visual Saliency Map to 1D Saliency Curves

In our task we need to transform the 2D saliency map to an 1D saliency curve. The simple spatial averaging across each frame is not suitable, since saliency maps contain many zero values in non-salient areas that affect the saliency curve. For this reason, we apply spatial averaging only on the salient regions of the saliency map. First, we define the operator $B : \mathbb{R}^{\mathbb{E}} \rightarrow \{0, 1\}^{\mathbb{E}}$ that transforms the saliency map into a binary image by applying the Otsu's threshold [49], where $\mathbb{E}$ denotes the image domain, i.e., a video frame of size $m \times n$. Then we take the 1D saliency curve $C(t)$ by applying the mapping $G : \mathbb{R}^{\mathbb{E}} \rightarrow \mathbb{R}$ on the saliency map $M(x, y, t)$:

$$C(t) = G(M(x, y, t)) = \frac{\sum_{x,y} B(M(x, y, t)) \cdot M(x, y, t)}{\sum_{x,y} B(M(x, y, t))} \tag{1}$$

Finally, we apply a median filter of length 151 frames to the saliency curve $C(t)$ and normalize its values in $[0, 1]$.

3.1.4 Fusion Schemes

For the fusion of the visual saliency components, we have experimented with two widely used functions: average and max, which correspond to different approaches in feature integration. Using the max, we search for regions or segments that are salient in at least one component, while by using the mean we need large saliency values in both components. Also, fusion can be applied at 2 different levels: a) the saliency map level or b) the saliency curve level after applying the transformation (1).
In the first case, the result remains a 2D saliency map $M_{ST}(x, y, t)$ from which the visual saliency curve $C_{ST}(t)$ is computed:

$$\text{Aver.}: \quad M_{ST}(x, y, t) = (M_S(x, y, t) + M_T(x, y, t))/2 \tag{2}$$

$$\text{Max}: \quad M_{ST}(x, y, t) = \max(M_S(x, y, t), M_T(x, y, t)) \tag{3}$$

$$C_{ST}(t) = G(M_{ST}(x, y, t)) \tag{4}$$

In the second case, fusion is performed between the obtained 1D saliency curves $C_S(t) = G(M_S(x, y, t))$, $C_T(t) = G(M_T(x, y, t))$:

$$\text{Aver.}: \quad \tilde{C}_{ST}(t) = (C_S(t) + C_T(t))/2 \tag{5}$$

$$\text{Max}: \quad \tilde{C}_{ST}(t) = \max(C_S(t), C_T(t)) \tag{6}$$

In Fig. 1a we see that the audio-visual saliency map is more correlated with the temporal visual component due to the enhancement from the audio saliency values. On the other case (Fig. 1b) the audio saliency curve modulates the temporal visual curve, which afterwards is fused with the static curve.

3.2. Auditory Saliency Model

Auditory saliency refers to the subset of the attention mechanisms that are responsible for the perception of sound information. The aim of the study of auditory saliency is to build a time-varying curve that resonates with the brain activation invoked to the listener of an the audio stream.

For audio saliency estimation, we employ Kayser et al. model [33], which is a behaviorally-inspired model and structurally identical to Itti et al. visual saliency model [29, 28], but has a different interpretation, as it integrates the concept of time. This model's input is a time-frequency representation of the signal, i.e., a spectrogram. The output is a saliency map, which depicts the evolution of auditory saliency over time and across frequencies. The extracted features are the intensity, temporal contrast, and frequency contrast, in various scales. Analogously to Itti et al. model, auditory saliency is estimated on the spectrogram image based on three low-level features: intensity, temporal contrast, and frequency contrast. As mentioned before, Kayser et al. model is behaviorally-inspired and thus, 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 such with an excitatory phase and a subsequent inhibitory one. These filters are modeled as Gabor filters with suitable orientations. A similar procedure of filtering and normalizing follows feature extraction and leads to a final 2-D saliency map.

3.3. Audio-Visual Fusion

In order to fuse the two modalities, which are inherently two non-comparable modalities with different dynamic ranges, we employ the following approach. First, regarding auditory saliency processing, as it was mentioned earlier, Kayser et al. model output is a 2D saliency map depicting saliency over time and frequencies. Since we are only interested in the time evolution of auditory saliency, we take the maximum saliency value on the map for each time instance, and thus obtain a 1D auditory saliency curve, denoted by SC_A. Subsequently, we exploit findings from neuroscience and relative behavioral experiments [64, 65, 9] that indicate that audiovisual integration is tolerant to an amount of asynchrony of maximum 200ms between audio and visual information. We integrate this finding by appropriately filtering auditory saliency via a Hanning window H of 200ms length centered at the current time instance [63]. Thus, we obtain a 1D curve that has incorporated this audiovisual temporal window of integration effect. The final auditory saliency curve is modeled as:

$$C_A(t) = \frac{1}{2N+1} \sum_{\ell=-N}^{N} SC_A(t+\ell) H(\ell),$$

where t is the video time index, ℓ is the audio sample index, and $2N+1$ the length of the window H.

The most important part of this model is the fusion between the auditory saliency curve and the visual saliency map. First, our fusion approach relies on the hypothesis that since audio features are dynamic/temporal, they influence only the dynamic/temporal visual saliency features. This hypothesis has also been verified through many experiments [18, 69, 56, 58, 47], such as the bouncing ball illusion [59]. Inspired by these findings, we fuse auditory saliency with temporal visual saliency. Specifically, fusion is applied between auditory saliency and each individual temporal feature of visual saliency separately. We combine these saliencies in a simple multiplicative manner, inspired by [52], where a similar approach has been followed. In their case, they deal with spatial audio, thus they have an auditory and a visual map with the same dimensions that combine by point-wise multiplication. In our case, audio is non-spatial, thus the resulting audio-temporal components for the 2D map and 1D curve representations are given by:

$$M_{TA}(x,y,t) = (1 + C_A(t)) \cdot M_T(x,y,t) \quad (7)$$

$$C_{TA}(t) = (1 + C_A(t)) \cdot C_T(t), \quad (8)$$

where $C_T(t) = G(M_T(x,y,t))$. After temporal-audio fusion, the spatial visual component is also integrated appropriately, according to the visual methods fusion strategy (Eqn. 2-6) in order obtain the audio-visual saliency map $M_{STA}(x,y,t)$ and curves $C_{STA}(t)$, $\tilde{C}_{STA}(t)$.

4. FMRI Movie Database

4.1. Experimental Design and Data Collection

For the fMRI data collection we decided to employ movie videos from COGNIMUSE database [73, 1]. The "COGNIMUSE database" is a new multimodal video dataset annotated with sensory and semantic saliency, events, cross-media semantics, and emotion. It can be used for training and evaluation of salient event detection and summarization algorithms, for classification and recognition of audio-visual and cross-media events, as well as for emotion tracking. Thus, the extension with fMRI data will be useful for many areas in computer vision and multimedia since researchers can take advantage of this additional

data in order to evaluate and design better computational models. Specifically, we have elected to present 20 minutes for each one of the five films: "Chicago" (CHI), "Crash" (CRA), "The Departed" (DEP), "Gladiator" (GLA), "Lord of the Rings - the Return of the King" (LOR) on the grounds that we have observed adequate discernible fluctuations in the corresponding annotated saliency curves. Each film excerpt was viewed by six different participants and the corresponding data has been collected.

The MRI images were acquired with a 3T Philips Achieva TX MRI scanner using gradient-echo EPI sequences (Time to Repetition – TR = 2 s, Field Of View – FOV of 192×240 mm^2, 36 sequential bottom-up transverse slices, voxel size 3×3×3 mm^3). Subjects were lying inside the scanner while the film excerpt was being back-projected on a semi-opaque material and they viewed the video through a mirror attached to the equipment. Headphones designated for usage inside MRI scanners were used for the audio stream.

4.2. fMRI Data Preprocessing

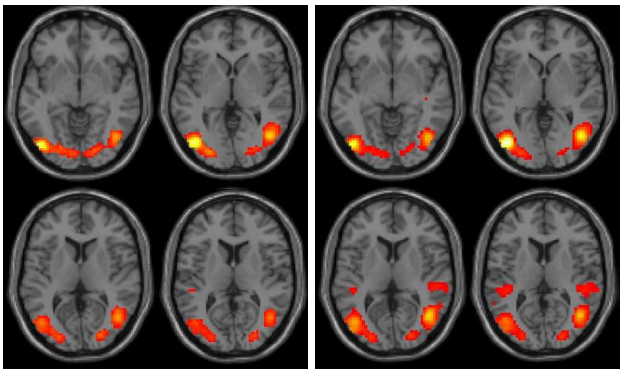
The SPM Toolbox [2] was used to preprocess the fMRI data and fit a General Linear Model (GLM). Raw data are spatially realigned (motion correction), temporally interpolated to compensate for acquisition delay, normalized to standard MNI space[1] and smoothed with an 8 mm wide Gaussian kernel. Following the preprocessing stage, high-pass filtering of 128 seconds cutoff is applied to the voxel time-series to remove low-pass physiological components such as respiration and heart beat. fMRI residual temporal autocorrelation was modeled as an autoregressive process AR(1) and integrated in the GLM estimation. Participants that had spontaneous movement above 4mm or 1 degree where excluded, unless transient movement could be removed by interpolation (scrubbing). We were thus left with 4, 6, 6, 5 and 5 participants for CHI, CRA, DEP, GLA and LOR respectively.

5. FMRI Analysis for Saliency Validation

5.1. Saliency Regressor Construction

As described in [51], in order to construct regressors suitable for the low-resolution fMRI time-series, based on saliency curves we need to sub-sample the curves from 25 values per second (per frame) to one value per 2 seconds (MRI scanner TR). The curves further need to be convolved with the standard haemodynamic response function (HRF), a low-pass function that introduces a time blurring and is considered to adequately model the transfer function of a voxel seen as a time-invariant linear system.

[1] Standard coordinate space for MRI data, based on the anatomical atlas by Montreal Institute of Neurology



(a) Visual Features (b) Audio-Visual Features

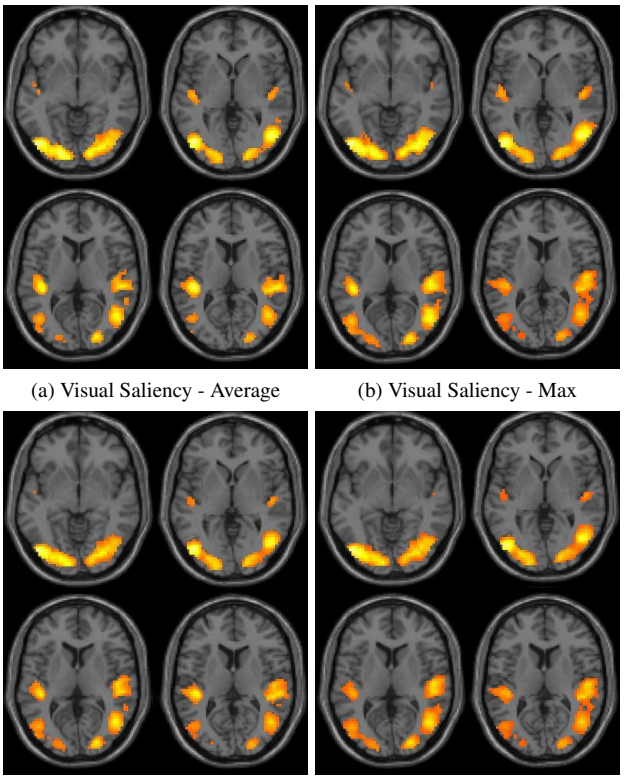
Figure 2: Results of GLM fit for visual and audio-visual *features* (F-test). Projection on transverse slices MNI z=-6, z=0, z=6 and z=10.

5.2. Mixed-effects Model for FMRI

In a manner similar to [51], the computationally constructed regressors are used to fit a General Linear Model (GLM) for each voxel independently. However, contrary to the previous works of [4, 51], which had been limited by the amount of data available to a fixed-effects approach, here we employ a mixed-effects model [48]. This approach is commonly accepted and widely used, since it accounts for variation across participants (inter-subject variation). It thus allows for generalization of results for the entire underlying population and is not restricted to the specific sample at hand. A mixed-effects model comprises of fixed-effect models fitted to the data of each participant individually (first level analysis) which are then combined in a random-effects group level model (second level analysis).

More specifically in SPM a "summary statistic" procedure is followed [17], whereby contrasts of effects of interest (in our case effects of regressors comprising features or saliency) are computed for each participant individually and then the corresponding statistical maps are used to fit an overall random effects model including all participants. Because T-contrasts are more reliable to take to the second level, when constructing regressors for the feature models, we manually orthogonalize features consecutively with respect to all the previous ones, so that $\mathbf{b}$ estimators will not be biased and we then use corresponding T-contrasts.

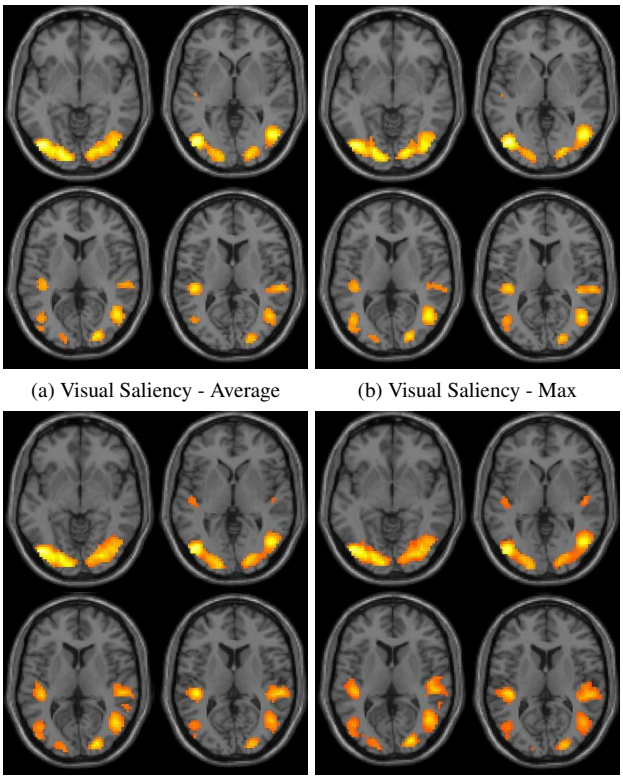
Also, in order to account for possible variation across movies that might induce a bias in the estimation, we also included in the group level model an extra regressor encoding the 5 different movies as a random effect. In the group level model, for feature models an F-contrast (based on F-statistics) was performed on $\mathbf{b}$ to test the overall variance of the observed data that could be explained by the model comprising the feature regressors. For saliency models, comprising only one regressor of interest, we use T-



(a) Visual Saliency - Average (b) Visual Saliency - Max

(c) Audio-Visual Saliency - Average (d) Audio-Visual Saliency - Max

Figure 3: Results of GLM fit for visual and audio-visual saliency with different fusion schemes in the level of *saliency maps* (t-test). Projection on transverse slices MNI z=-6, z=0, z=6 and z=10.



(a) Visual Saliency - Average (b) Visual Saliency - Max

(c) Audio-Visual Saliency - Average (d) Audio-Visual Saliency - Max

Figure 4: Results of GLM fit for visual and audio-visual saliency with different fusion schemes in the level of *saliency curves* (t-test). Projection on transverse slices MNI z=-6, z=0, z=6 and z=10.

contrasts, which have the additional advantage of being directional (i.e. differientiate between positive and negative correlations) in contrast to F-contrasts which only measure the amount of variance explained. For voxels whose p-value satisfies the p-FWE = 0.05 threshold, corrected for multiple comparisons (family-wise error correction), the model regressors associated with the F- or T-contrast are considered to have a good predictability of the voxel time-series [16].

6. Evaluation

Results are presented in the form of thresholded statistical maps [51] that are produced by the GLM analysis. In the figures that follow, we have elected to present axial slices with z=-6, z=0, z=6 and z=10 (MNI space coordinates), which allow for a good and concise view of both the visual and the auditory cortex and can facilitate a comparison between the visual and the audiovisual saliency (or feature) models. The color scale runs from red to white, the latter corresponding to the highest predictability.

6.1. Evaluation of Visual and Auditory Features

In Fig. 2 we present the results regarding the 3 different features curves (static visual saliency, temporal visual saliency, auditory saliency) which have the expected pattern. The works of [4, 51] have pointed the brain areas that are activated when a visual or audio stimuli is attended. More specifically, when only visual features are employed we are able to predict only the responses of voxels in the visual cortex while adding the auditory feature activation in the auditory cortex is also present. With this analysis we were able to validate the appropriateness of these three independent subsystems for the task of temporal saliency prediction.

6.2. Evaluation of Audio-Visual Model for Temporal Saliency

Afterwards, we proceed to the evaluation of the proposed audio-visual temporal saliency model. In order to validate the proposed fusion and integration schemes we expect to have the same behavior as in the previous analysis by em-

Table 1: Location of activation peaks for *visual* saliency model (fusion in curves level with max function) [2].

MNI Coordinates X Y Z	*Hemisphere*	*Functional or [Anatomical] area*	*T-value*
-42 -73 -2	Left	V5/MT	14.23
-18 -85 -8	Left	V3 ventral	12.69
39 -73 -11	Right	V4 ventral	11.58
45 -67 1	Right	V5/MT	11.01
-33 -82 -8	Left	V4 dorsal	10.98
-42 -28 10	Left	Area TE 1.1	8.69
-39 -31 13	Left	Area TE 1.1	8.67

Table 2: Location of activation peaks for *audio-visual* saliency model (fusion in curves level with max function).

MNI Coordinates X Y Z	*Hemisphere*	*Functional or [Anatomical] area*	*T-value*
-45 -76 -2	Left	V4 ventral	16.69
-21 -82 -8	Left	V3 ventral	12.22
48 -67 1	Right	V4 ventral	12.07
36 -73 -8	Right	V5/MT	11.72
-9 -91 -5	Left	V1	10.87
-39 -31 13	Left	Area TE 1.1	10.70
-48 -31 16	Left	Area TE 1.1	10.67
-42 -28 10	Left	Area TE 1.1	10.43

ploying only one curve rather than independent features series. In Figs 3,4 we present the results regarding the visual and audio-visual models for the two different fusion levels respectively.

Looking at the figures, we can clearly see that although the results in the visual cortex are almost identical between the visual and the audiovisual model, the latter also has visible clusters of activation in the auditory cortex, especially where we apply fusion at the curve level. Traces of activation within the auditory cortex can be found in the visual saliency model as well, which are, however, limited both in extent and peak magnitude. We should also bear in mind that the visual and auditory modality in natural stimuli (i.e. not deliberately manipulated) are often correlated in the first place, which has an innate effect on our results.

Regarding the level where fusion is performed, it seems that fusion at the curve level works better for the task of temporal saliency prediction than fusion of the saliency maps which are extensively employed in the task of fixation prediction. In addition, when we compare the two fusion schemes we see that the nonlinear fusion with max works slightly better than the average, since it gives more focused activation in the visual and auditory cortex areas.

In Tables 1,2 we present the locations of the top activation peaks for the visual and audio-visual saliency models respectively, for fusion at the curve level with max function. Comparing the two tables, we can observe that they are quite similar. Also, peaks within the auditory cortex can be found both in the audio-visual as well as in the visual model. However, one can notice the difference in the T-statistic value, which is much higher for the audio-visual model.

Table 3: Visual vs. audio-visual saliency model: % of voxels of each visual or auditory area that shows significant association (fusion in curves level with max function) [3].

Visual area	*% for Visual Model*	*% for AV Model*
V1	5.90	7.50
V2	1.40	1.80
V3 ventral	12.10	13.60
V3 dorsal	0.25	1.00
V4 ventral	23.40	24.30
V4 dorsal	32.10	35.80
V5/MT	96.50	98.60
Auditory area	*% for Visual Model*	*% for AV Model*
TE 1.0	19.2 (L only)	63.25
TE 1.1	51.5	76.40
TE 1.2	0	4.6 (L only)
TE 3	1.00 (R only)	1.2 (R only)

Table 3 depicts the voxel percentage of each visual or auditory area that shows significant association, which can be considered as a measure for the overall sensitivity of the proposed saliency models. We highlight the fact that the overall detection is significant higher for the auditory areas when the audio-visual model is employed, while detection in the visual areas remains high for both models.

7. Conclusion

In this work we proposed an audio-visual approach for tracking the temporal saliency in video as well as an alternative way for validation using fMRI data. We developed a computational audio-visual saliency model by employing deep learning architectures that are originally proposed for the task of fixation prediction, where we additionally incorporated several behavioral findings related to audiovisual integration. In addition, we collected a fMRI database, which contains both different videos and subjects, that may become useful for many computer vision problems related to the video domain. The fMRI-based evaluation showed that the proposed audio-visual model has high correlation with the brain responses and both the visual and auditory cortex are localized. As future work, we intend to extend the audio-visual saliency model by designing a deep architecture that will be trained jointly by employing audio-visual data. We also plan to take advantage of the fMRI data in order to find feature embeddings that will augment the current deep methods with the human brain information.

Acknowledgements: This work was partially supported by EU projects Babyrobot/687831 and i-Support/643666.

[2] Brain areas are given in conventional notation. See Anatomy Toolbox atlas for full names.

[3] Unless otherwise specified average of left and right hemisphere is given, since no laterality effect was observed.

References

[1] COGNIMUSE Project. http://cognimuse.cs.ntua.gr/. 5

[2] SPM Toolbox Software – University College London. http://www.fil.ion.ucl.ac.uk/spm/software/spm12/. 6

[3] C. Bak, A. Kocak, E. Erdem, and A. Erdem. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Trans. on Multimedia*, 2017. 3

[4] C. Bordier, F. Puja, and E. Macaluso. Sensory processing during viewing of cinematographic material: Computational modeling and functional neuroimaging. *NeuroImage*, 67(2):213226, 2013. 2, 3, 6, 7

[5] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(1):185–207, 2013. 2

[6] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Trans. Image Processing*, 22(1):55–69, 2013. 1, 2

[7] N. D. B. Bruce and J. K. Tsotsos. Spatiotemporal saliency: Towards a hierarchical representation of visual saliency. In *Int. Workshop on Attention and Performance in Comp. Vis.*, 2008. 2

[8] M. K. Carroll, G. A. Cecchi, I. Rish, R. Garg, and A. R. Rao. Prediction and interpretation of distributed neural activity with sparse models. *NeuroImage*, 44(1):112–122, 2009. 3

[9] L. Chen and J. Vroomen. Intersensory binding across space and time: A tutorial review. *Attention, Perception, & Psychophysics*, 75(5):790–811, 2013. 5

[10] A. Coutrot and N. Guyader. An audiovisual attention model for natural conversation scenes. In *Proc. IEEE Int. Conf. on Image Processing*, pages 1100–1104, 2014. 3

[11] A. Coutrot and N. Guyader. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision*, 14(8):1–17, 2014. 3

[12] I. Daubechies, E. Roussos, S. Takerkart, M. Benharrosh, C. Golden, K. D'ardenne, W. Richter, J. D. Cohen, and J. Haxby. Independent component analysis for brain fMRI does not select for independence. *Proc. National Academy of Sciences*, 106(26):10415–10422, 2009. 3

[13] N. Ejaz, I. Mehmood, and S. W. Baik. Efficient visual attention based framework for extracting key frames from videos. *Signal Processing: Image Communication*, 28(1):34–44, 2013. 1

[14] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, textual attention. *IEEE Trans. Multimedia*, 15(7):1553–1568, 2013. 1, 3

[15] G. Evangelopoulos, A. Zlatintsi, G. Skoumas, K. Rapantzikos, A. Potamianos, P. Maragos, and Y. Avrithis. Video event detection and summarization using audio, visual and text saliency. In *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, pages 3553–3556, 2009. 3

[16] K. J. Friston, A. P. Holmes, K. J. Worsley, J. Poline, C. D. Frith, and R. S. J. Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, 2(4):189–210, 1994. 7

[17] K. J. Friston, K. E. Stephan, T. E. Lund, A. Morcom, and S. Kiebel. Mixed-effects and fMRI studies. *Neuroimage*, 24(1):244–252, 2005. 6

[18] J. W. Gebhard and G. H. Mowbray. On discriminating the rate of visual flicker and auditory flutter. *The American Journal of psychology*, 72(4):521–529, 1959. 5

[19] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008. 3

[20] C. Guo and L. Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. Image Processing*, 19(1):185–198, 2010. 3

[21] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems (NIPS)*, 2006. 2

[22] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303:1634–1640, 2004. 3

[23] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. In *Proc. IEEE Int. Conf. on Computer Vision*, 2017. 1

[24] X. Hou and L. Zhang. Dynamic visual attention: searching for coding length increments. In *Advances in Neural Information Processing Systems (NIPS)*, 2009. 2

[25] X. Hu, F. Deng, K. Li, T. Zhang, H. Chen, X. Jiang, J. Lv, D. Zhu, C. Faraco, and D. Zhang. Bridging low-level features and high-level semantics via fMRI brain imaging for video classification. In *Proc. 18th ACM Int. Conf. Multimedia*, 2010. 3

[26] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proc. IEEE Int. Conf. on Computer Vision*, 2015. 3

[27] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2005. 2

[28] L. Itti, N. Dhavale, and F. Pighin. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proc. 48th SPIE Int. Symp. Optical Science and Technology*, volume 5200, 2003. 2, 5

[29] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20:1254–1259, 1998. 3, 5

[30] S. Jetley, N. Murray, and E. Vig. End-to-end saliency mapping via probability distribution prediction. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016. 3

[31] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. IEEE, 2015. 1

[32] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convo-

lutional neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. 1

[33] C. Kayser, C. Petkov, M. Lippert, and N. Logothetis. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, 15:1943–1947, 2005. 5

[34] I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2017. 1

[35] P. Koutras and P. Maragos. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication*, 38:15–31, 2015. 2

[36] P. Koutras, A. Zlatintsi, E. Iosif, A. Katsamanis, P. Maragos, and 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*, pages 4361–4365, 2015. 1, 3

[37] M. Kmmerer, L. Theis, and M. Bethge. Deep Gaze I: Boosting saliency prediction with feature maps trained on imagenet. In *Proc. Int. Conf. of Learning Representations (ICLR) Workshop*, 2015. 3

[38] G. Leifman, D. Rudoy, T. Swedish, E. Bayro-Corrochano, and R. Raskar. Learning gaze transitions from depth to improve video saliency estimation. In *Proc. IEEE Int. Conf. on Computer Vision*, 2017. 3

[39] T. Liu, X. Hu, X. Li, M. Chen, J. Han, and L. Guo. Merging neuroimaging and multimedia: methods, opportunities and challenges. *IEEE Trans. Human-Machine Systems*, 44(2), 2014. 2

[40] J. Lv, X. Jiang, X. Li, D. Zhu, H. Chen, T. Zhang, S. Zhang, X. Hu, J. Han, and H. Huang. Sparse representation of whole-brain fMRI signals for identification of functional networks. *Medical Image Analysis*, 20(1):112–134, 2015. 3

[41] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(1):171–177, 2010. 2

[42] P. Maragos, A. Gros, A. Katsamanis, and G. Papandreou. Cross-modal integration for performance improving in multimedia: A review. In *Multimodal Processing and Interaction: Audio, Video, Text*, pages 1–46. Springer-Verlag, 2008. 1

[43] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976. 3

[44] M. A. Meredith and B. E. Stein. Interactions among converging sensory inputs in the superior colliculus. *Science*, 221(4608):389–391, 1983. 1

[45] M. A. Meredith and B. E. Stein. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology*, 56(3):640–662, 1986. 1

[46] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1-2):145–175, 2004. 3

[47] S. Morein-Zamir, S. Soto-Faraco, and A. Kingstone. Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, 17(1):154–163, 2003. 5

[48] J. A. Mumford and R. A. Poldrack. Modeling group fMRI data. *Social cognitive and affective neuroscience*, 2(3):251–257, 2007. 6

[49] N. Otsu. A threshold selection method from gray-level histograms. *Trans. on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. 4

[50] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016. 3, 4

[51] G. Panagiotaropoulou, P. Koutras, A. Katsamanis, P. Maragos, A. Zlatintsi, A. Protopapas, E. Karavasilis, and N. Smyrnis. fMRI-based perceptual validation of a computational model for visual and auditory saliency in videos. In *Proc. Int. Conf. Image Processing*, 2016. 2, 6, 7

[52] S. Ramenahalli, D. R. Mendat, S. Dura-Bernal, E. Culurciello, E. Nieburt, and A. Andreou. Audio-visual saliency map: Overview, basic models and hardware implementation. In *Proc. Information Sciences and Systems (CISS)*, pages 1–6, 2013. 3, 5

[53] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2013. 4

[54] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer. Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. In *Int. Conf. on Robotics and Automation*, pages 962–967, 2008. 3

[55] B. Schauerte, B. Kühn, K. Kroschel, and R. Stiefelhagen. Multimodal saliency-based attention for object-based scene analysis. In *IEEE Int. Conf. on Intelligent Robots and Systems*, pages 1173–1179. IEEE, 2011. 3

[56] R. Sekuler, A. B. Sekuler, and R. Lau. Sound alters visual motion perception. *Nature*, 385(6614):308, 1997. 5

[57] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15, 2009. 2

[58] L. Shams, Y. Kamitani, and S. Shimojo. What you see is what you hear. *Nature*, 408:788, 2000. 5

[59] S. Shimojo and L. Shams. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11(4):505–509, 2001. 3, 5

[60] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. of Learning Representations (ICLR)*, 2015. 1

[61] G. Song. *Effect of sound in videos on gaze: Contribution to audio-visual saliency modeling*. PhD thesis, Universite de Grenoble, 2013. 3

[62] A. Toet. Computational versus Psychophysical Image Saliency: A Comparative Evaluation Study. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(11), 2011. 2

[63] A. Tsiami, A. Katsamanis, P. Maragos, and A. Vatakis. Towards a behaviorally-validated computational audiovisual saliency model. In *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, pages 2847–2851, 2016. 2, 3, 5

[64] E. Van der Burg, C. N. L. Olivers, A. W. Bronkhorst, and J. Theeuwes. Pip and pop: Nonspatial auditory signals im-

prove spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34:1053–1065, 2008. 5

[65] V. van Wassenhove, K. W. Grant, and D. Poeppel. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3):598–607, 2007. 5

[66] A. Vatakis and C. Spence. Crossmodal binding: Evaluating the unity assumption using audiovisual speech stimuli. *Perception & Psychophysics*, 69(5):744–756, 2007. 1

[67] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. European Conf. on Computer Vision*, 2016. 4

[68] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji. Revisiting video saliency: A large-scale benchmark and a new model. 2018. 1, 4

[69] R. B. Welch, L. D. DutionHurt, and D. H. Warren. Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics*, 39(4):294–300, 1986. 5

[70] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognition Symposium*, pages 214–223. Springer, 2007. 4

[71] L. Zhang, M. H. Tong, and G. W. Sunday: Saliency using natural statistics for dynamic analysis of scenes. In *Proc. Thirty-first Annual Cognitive Science Society Conference.*, 2009. 2

[72] S. Zhao, X. Jiang, J. Han, X. Hu, D. Zhu, J. Lv, T. Zhang, L. Guo, and T. Liu. Decoding auditory saliency from fMRI brain imaging. In *Proc. 22nd ACM Int. Conf. Multimedia*, 2014. 3

[73] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, and P. Maragos. COGNIMUSE: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017(1):54, Aug 2017. 5