Proceedings of IEEE Int'l Conference on Image Processing (ICIP-2016), Phoenix, AZ, USA, Sep. 2016.

FMRI-BASED PERCEPTUAL VALIDATION OF A COMPUTATIONAL MODEL FOR VISUAL AND AUDITORY SALIENCY IN VIDEOS

G. Panagiotaropoulou[1], *P. Koutras*[1,5], *A. Katsamanis*[1,5], *P. Maragos*[1,5], *A. Zlatintsi*[1], *A. Protopapas*[2], *E. Karavasilis*[3], *N. Smyrnis*[4]

[1]School of Electr. and Comp. Engineering, National Technical Univ. of Athens, 15773 Athens, Greece

[2]Department of Philosophy and History of Science, Univ. of Athens, 15771 Zografos, Greece

[3]Radiology and Medical Imaging Research Unit, Univ. of Athens, Greece

[4]Department of Psychiatry, Eginition Hospital, School of Medicine, Univ. of Athens, Greece

[5]Athena Research and Innovation Center, Maroussi 15125, Greece

gio.panagiotaropoulou@gmail.com, {pkoutras, nkatsam, maragos, nzlat}@cs.ntua.gr

aprotopapas@phs.uoa.gr

ABSTRACT

In this study, we make use of brain activation data to investigate the perceptual plausibility of a visual and an auditory model for visual and auditory saliency in video processing. These models have already been successfully employed in a number of applications. In addition, we experiment with parameters, modifications and suitable fusion schemes. As part of this work, fMRI data from complex video stimuli were collected, on which we base our analysis and results. The core part of the analysis involves the use of well-established methods for the manipulation of fMRI data and the examination of variability across brain responses of different individuals. Our results indicate a success in confirming the value of these saliency models in terms of perceptual plausibility.

Index Terms— visual saliency, auditory saliency, fMRI, General Linear Model, spatio-temporal Gabor energy filterbank, AM-FM sound analysis

1. INTRODUCTION

As a cognitive mechanism employed by both humans and artificial systems, attention has been an active research field for many years now. Selecting the most important part of information for further processing, attention mechanisms play a crucial role in human perception and thus have numerous applications. The feature integration theory [2] and the concept of spatial saliency maps [3] became the basis of many cognition-inspired attention models, such as [4], [5] or [6] for vision and [7] or [8] for audition. In the context of video processing, an audiovisual modeling of human attention has been attempted in [9].

Significant neurobiological and psychophysical evidence indicates that the first stages of sensory information processing include many feature detection processes. Brain imaging techniques, such as functional Magnetic Resonance Imaging (fMRI), in conjunction with computational methods 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 improve computational schemes for feature extraction [10].

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 [11], [12], [13], at assessing the biological plausibility of widespread perceptual models [14], [15] or at augmenting the latter by integrating high-level information encoded inside the human brain [16], [17]. 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 [18].

In this study, we test the validity of a perceptually-inspired visual model [19] and we investigate the biological plausibility of a previously proposed auditory saliency framework [20]. These models have been successfully employed in a number of applications (see section 2). In addition, we experiment with parameters, modifications and fusion schemes that render these models more fit to describe brain activation data and thus more compatible with the functions of the human brain. Since fMRI data are scarce, especially when it comes to naturalistic stimuli, we opted to experiment on data newly collected as part of this study, rather than to rely on already existing databases. For our purposes we employed a design similar to [14], which we first validate on our setup. More specifically, we use the visual and auditory features as regressors to reconstruct the time-series of voxels in the brain by means of a General Linear Model (GLM), a method well-established for use with fMRI. We expect to obtain an accurate reconstruction for voxels in the visual and auditory cortex respectively. We then perform a cross-check on our results based on the correlation of brain areas responses across subjects. Our results indicate a success in verifying the perceptual validity of the computational models in both modalities.

First, we briefly describe the visual and auditory front-ends (section 2). We next present our experimental setup for data collection

This work was partially supported by the project "COGNIMUSE" [1], under the "ARISTEIA" Action, co-funded by the European Social Fund and Greek National Resources. It was also partially supported by the European Union under the projects BabyRobot with grant H2020-687831 and DIRHA with grant FP7-288121.

(section 3), followed by the method used to fit a model on the features and fMRI data in section 4. Finally, in section 5 we present and evaluate our results.

2. COMPUTATIONAL SALIENCY MODELS

2.1. Visual model

In this work we employ a variant of a recently proposed spatio-temporal model for visual saliency that has achieved a good performance in many application tasks such as eye-fixation prediction, action classification and movie summarization [21, 19]. Specifically, we employed a spatio-temporal filterbank of 400 3D Gabor filters [21], as described in [19] for both the luminance and color streams. This frontend provides motion information in different scales and directions, thus detecting both the fastest changes in the video stimuli (e.g. flicker) and the slowest motion changes related to action events. Instead of keeping only the dominant energy, we extract frame-wise features that correspond to 7 dominant spatio-temporal energies in order to increase robustness. We then apply a simple fusion scheme by normalizing each feature time-series and then taking the mean over all features. In this way, we obtain a single value describing the saliency of each video frame.

2.2. Auditory model

For the auditory modeling we employ the audio features proposed in [20]. These features are based on AM-FM demodulation [22] and variants have been successfully used in many applications such as speech and music recognition [22] and summarization [9, 20]. The employed feature vectors consist of 25 Teager-Kaiser energies that are extracted using a Gabor filterbank. Sound loudness and roughness, which have been found to correlate with the functioning of the human auditory system, are included as additional features.

We used a variant of the proposed scheme, where only 12 Teager-Kaiser energies were used to avoid estimability issues with the GLM: due to low time resolution of the fMRI, available data points suffice for a reliable estimation of a limited amount of parameters. The logarithm of each narrow-band signal energy was then taken before fusing the energy features to obtain the final saliency curve using a max operation, since we have found the resulting curve to be more descriptive of a low-pass physiological signal, such as BOLD. This aimed at suppressing sharp peaks produced by the Teager-Kaiser energy operator, which have proven to be of value to a machine learning based summarization scheme [21] nonetheless.

3. EXPERIMENTAL DESIGN AND SETUP

3.1. Experimental design

Validation design. In order to appraise the validity of our setup, we tested the proposed scheme using a paradigm that permits straightforward interpretation of the results. Thus, the visual and auditory stream had been independently manipulated so that they formed segments with and without sound and segments with color image, grayscale image and no image at all (ON/OFF design). This way, the two streams are also artificially de-correlated. For this purpose we used an excerpt of a wildlife documentary. Data from 5 participants were collected.

Saliency extraction design. We will further test to what extent the desired results can be reproduced for normal free-viewing conditions without any manipulation. In our implementation we have elected to present the first 20 minutes of the feature film "The Departed" from the COGNIMUSE [1], [9] database of annotated films, on the grounds that we have observed adequate discernible fluctuations in the corresponding saliency curve. Data from 6 participants have been collected.

3.2. fMRI data collection and pre-processing

The 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. The SPM Toolbox [23] was used to preprocess the fMRI data and fit a 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 AR(1) process and integrated in the GLM estimation (see Section 4.2).

4. THE GENERAL LINEAR MODEL FOR FMRI DATA

4.1. Regressor construction

Visual as well as auditory features, and therefore saliency curves produced via fusion are provided by the selected front-ends on a frame basis, that is one value per frame or 25 values per second. In order to use them as regressors for the low-resolution fMRI time-series, subsampling to one value per 2 seconds is required to match the MRI scanner TR. Following subsampling, regressors are convolved with the standard haemodynamic response function (HRF). This low-pass function introduces a time blurring and is considered to adequately model the transfer function of a voxel seen as a time-invariant linear system. The sets of visual and auditory regressors are used to fit two models independently for each modality. Likewise, the visual and auditory saliency regressors are used to fit two independent models.

4.2. Model description and estimation

The computationally constructed features are used to fit a GLM for each voxel independently. fMRI time-series are thus represented as $\mathbf{y} = \mathbf{X}\mathbf{b} + \epsilon$, where N is the number of volumes (one per TR), $\mathbf{y}$ is an $N \times 1$ vector comprising the observed fMRI voxel time-series (predictant variable), $\mathbf{X}$ is the $N \times K$ design matrix comprising the regressors (predictor variables) that have been computed from the visual or auditory features (one regressor per column), as well as the motion correction regressors estimated during preprocessing and a constant regressor modeling baseline activation, $\mathbf{b}$ is a $1 \times N$ vector representing the weights (beta values) attributed to each regressor according to the model estimation and ϵ is an additive Gaussian error. The model parameters, including hyper-parameters modeling temporal autocorrelation are jointly estimated using a Restricted Maximum Likelihood algorithm (ReML) [24], [25]. Hence, $\mathbf{b}$ are voxel-specific, whereas $\mathbf{X}$ is identical for all voxels.

We have employed a fixed-effects approach, using data from each subject as a different session in the model by concatenating

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

the data vertically and using a separate constant regressor for each subject. The assumption upon which this setup is based, namely that the model underlying the data is identical across subjects, is a bold one to make for natural data (here brain activity). However, the small sample (n=6 participants) does not allow for random-effects analysis that would permit more robust generalizations [26].

Regarding 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. This way we can test the joint contribution of all regressors, despite high cross-correlations between them. Such correlations render the use of a T-contrast (based on T-statistics) to assess each regressor independently essentially meaningless. For saliency models, comprising only one regressor of interest, we use T-contrasts. 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 [25].

5. EXPERIMENTAL RESULTS

Results are presented in the form of thresholded statistical maps. The color scale runs from red to white, the latter corresponding to the highest predictability. Additionally, activation peaks and their locations are reported, as well as a summary of the model's ability to detect voxels from the corresponding modality while avoiding false positives. The anatomical and functional labeling of activated clusters and peaks was performed according to the atlas provided by the Anatomy Toolbox. Whenever no assignment to a functional area could be found, the anatomical area is reported.

5.1. Results for the validation design

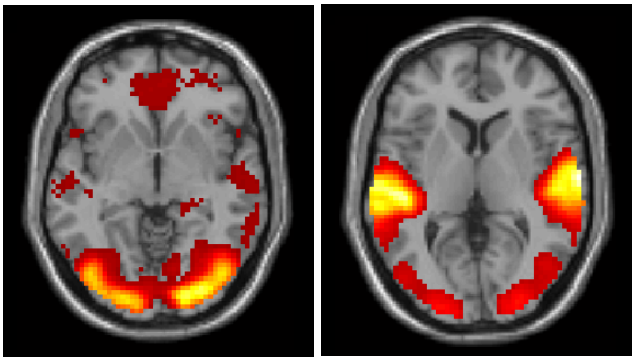
As can be seen from Fig. 1, the results we obtained have the expected pattern. More specifically, the visual model is able to predict the responses of voxels in the visual cortex quite accurately, while few voxels have been detected outside of that. Some of the activation that is not specific to the visual cortex could be attributed to multimedia content related processing, such as areas involved in working memory.

Likewise, the auditory model has a very good localization of the auditory cortex. The observed activation in the visual cortex is actually indicative of an anti-correlation between the auditory regressors and the visual cortex activity; intense stimulus in a modality is known to often have a inhibitory effect on the complementary one [27]. We have tested this via a directional t-test in the auditory saliency model: no activation, that is, no positive relation, was found with voxels in the visual cortex.

5.2. Results for visual features and saliency

We now turn to the results for the feature film. Figure 2 displays four representative transverse slices of thresholded maps in the MNI space, both for visual features and visual saliency models. We can observe that the pattern of activation is quite satisfactory, although in the case of the feature model there is some evident cross-modal activation; activated voxels extend to part of the auditory cortex and other areas. However, as is evident from Table 1 and Table 2 both the peaks and the overall detection are well localized. The saliency model as well produces satisfactory activation peaks (Table 3).

We observe very high activation in area V5, encoding motion [28] and other areas high in the hierarchy of the visual cortex, while restricted activation and absence of peaks is noted for the primary (V1) and secondary (V2) visual cortex (see section 5.4 for further details). Co-activation of the auditory cortex may be partly attributed to the correlation between modalities, namely to the fact that visually salient events coincide with acoustic events that are also perceived as salient.



(a) Visual features (z=-4) (b) Auditory features (z=5)

Fig. 1. Results of GLM fit for visual (a) and auditory (b) features. Transverse slices in MNI space.

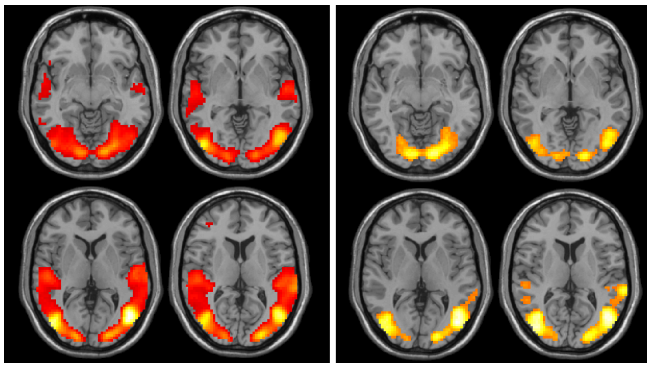


Fig. 2. Results of GLM fit for visual features (left – F-test) and saliency (right – t-test). Projection on transverse slices MNI z=-8 to z=10 with step 8.

Table 1. Location of activation peaks for visual features model [2].

MNI Coordinates X Y Z	*Hemisphere*	*Functional or [Anatomical] area*	*F-value*
45 -67 4	Right	V5/MT	36.23
-48 -76 4	Left	V5/MT	30.47
27 -91 16	Right	V3	21.01
27 -79 -11	Right	V4	18.59
-21 -91 7	Left	V3	15.43

Table 2. Visual features model: % of voxels of each visual area that show significant association.

Visual area	*% in Left Hemisphere*	*% in Right Hemisphere*
V1	11.00	18.60
V2	4.60	16.70
V3 ventral	33.30	42.20
V3 dorsal	3.80	16.50
V4 ventral	66.60	55.30
V4 dorsal	36.00	36.60
V5/MT	100.00	94.60

In order to test the robustness of this visual front-end, we also experimented with using only two dominant energies as regressors – one for luminance and one for color. This produced results with a

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

Table 3. Location of activation peaks for visual saliency model.

MNI Coordinates X Y Z	*Hemisphere*	*Functional or [Anatomical] area*	*T-value*
42 -73 4	Right	[Mid. Occip. Gyrus]	11.35
-42 -64 4	Right	[Mid. Temp. Gyrus]	11.00
27 -91 16	Right	V5/MT	10.55
27 -79 -11	Right	[Mid. Temp. Gyrus]	10.16
-21 -91 7	Left	V4	10.15

similar, yet more shrunk pattern localized around area V5 bilaterally. Such result is expected, since spatio-temporal energies are capable of modeling both slow and rapid motion changes in a video [21].

5.3. Results for auditory features and saliency

As far as auditory models are concerned, we obtain results with strong presence of activation in the auditory cortex both for the case of features and saliency. Nevertheless, cross-modal effects are far more evident here. It seems that concomitant visual events with conspicuous motion introduce significant correlation between auditory feature (and saliency) regressors and brain voxels located in the visual cortex, since motion is explicitly encoded in the V5 [28]. Activation peaks (Tables 4 and 5) support this hypothesis, since they are located in the auditory cortex and area V5.

By examination of the auditory regressors and motion estimation, we found that large peaks describing or coinciding with motion dominate parameter estimation in the GLM and thus result in visual voxels scoring above the statistical activation threshold. This cross-modal effect is also true for visual features, though to a lesser extent.

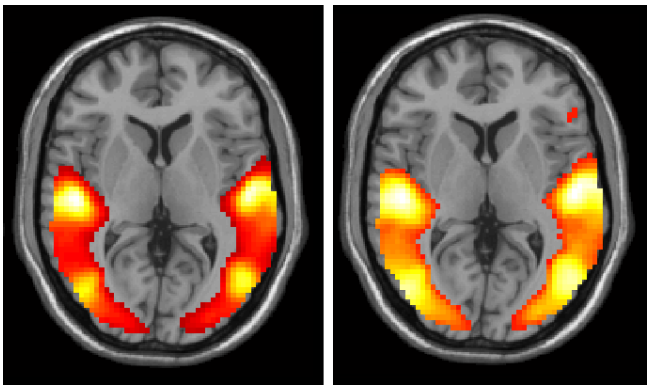


Fig. 3. Results of GLM fit for auditory features (left – F-test) and saliency (right – t-test). Projection on transverse slice MNI z=5.

Table 4. Location of activation peaks for auditory features model.

MNI Coordinates X Y Z	*Hemisphere*	*Functional or [Anatomical] area*	*F-value*
-5 -19 4	Left	TE 1.0	52.16
54 -13 4	Right	TE 1.0	49.04
48 -70 4	Right	V5/MT	38.56
-39 -31 13	Left	TE 1.1	36.05
-48 -73 7	Left	V5/MT	35.85

5.4. Study of brain response correlation across subjects

In an effort to further establish the validity of our results, we conducted an analysis based on inter-subject correlation. We extracted the mean response of each anatomically defined brain area and calculated the pairwise Pearson correlation coefficient between all subjects. Then we performed a t-test to check whether each mean between-subjects coefficient differed significantly from zero.

Table 5. Location of activation peaks for auditory saliency model.

MNI Coordinates X Y Z	*Hemisphere*	*Functional or [Anatomical] area*	*T-value*
Cluster 1 (Left)			
51 -22 7	Left	TE 1.0	20.34
-48 -76 4	Left	V5/MT	17.72
-39 -31 13	Left	TE 1.1	16.84
Cluster 2 (Right)			
51 -19 7	Right	TE 1.0	20.14
48 -70 4	Right	V5/MT	17.86
66 -25 10	Right	[Sup. Temp. Gyrus]	16.83

As we can see in Table 6, areas with high inter-subject correlation have been detected by the visual and auditory models respectively. It is also worth noting that V5 encoding motion indeed displays the highest value. We have also reported a small participation of areas V1 and V2, both low in the visual system hierarchy. We can derive from Table 6, that this in not only due to the model's incapability to provide a plausible representation of the information encoded in these areas, but also due to the discordance among subjects. Part of this effect can be attributed to the experimental setup and the projection technique in particular: the semi-opaque material blurs the image and renders low-level static information primarily encoded in these areas, such as edges, less conspicuous.

It is also worth noting that this analysis produced similar results for the ON/OFF validation design.

Table 6. Auditory and visual brain areas with statistically significant ($p < 0.001$) inter-subject correlation in order of ascending p-values (column-wise).

Funct. Area	*cont'd*	*cont'd*	*cont'd*
V5/MT L	TE 1.0 L	TE 1.2 L	TE 1.2 R
V5/MT R	FG1 R	V4 dorsal R	V1 L
V4 ventral R	TE 1.1 R	V3 ventral R	V2 R
TE 1.0 R	V4 dorsal L	FG2 R	V1 R
FG1 L	Area TE 3 L	V3 dorsal R	V2 L
V4 ventral L	V3 ventral L	V3 dorsal L	
Area TE 1.1 L	FG2 L	TE 3 R	

6. CONCLUSIONS

In this study we tested the perceptual validity of computational models for visual and auditory saliency, both of which have been successfully used in a number of applications, using fMRI data. Imaging data were collected for an excerpt of a feature film as part of this work. To this end we employed widely accepted methods, such as GLM modeling of fMRI voxel responses and correlation of mean responses of anatomically defined brain areas across subjects. We have achieved satisfactory results for vision, thus experimentally establishing the proposed model as a perceptually inspired one. As pertains to audition, we have obtained promising results, which attest to the suitability of the model in dealing with problems of human attention in multimodal video processing.

Our current efforts aim at extending our fMRI database to more video stimuli and more participants, which will broaden the scope and enhance the validity of our arguments. We are also investigating more elegant and effective fusion schemes, as well as more sophisticated methods for linking our computational models to brain imaging data, such as ones that make use of temporal information.

7. ACKNOWLEDGEMENTS

The authors wish to thank all the members of the CVSP Lab and the students of the Interdepartmental Program in Basic and Applied Cognitive Science who volunteered for fMRI data collection.

8. REFERENCES

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

[2] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[3] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

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

[5] 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'l Symp. Optical Science and Technology*, 2004, pp. 64–78.

[6] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2006, pp. 545–552.

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

[8] E. M. Kaya and M. Elhilali, "A temporal saliency map for modeling auditory attention," in *Proc. 46th Annual Conf. Information Sciences and Systems (CISS)*, 2012.

[9] 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, and textual attention," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[10] 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*, vol. 44, no. 2, pp. 270–280, 2014.

[11] 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*, vol. 44, no. 1, pp. 112–122, 2009.

[12] 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*, vol. 57, no. 1-2, pp. 145–175, 2004.

[13] 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*, vol. 106, no. 26, pp. 10415–10422, 2009.

[14] C. Bordier, F. Puja, and E. Macaluso, "Sensory processing during viewing of cinematographic material: Computational modeling and functional neuroimaging," *NeuroImage*, vol. 67, no. 2, pp. 213–226, 2013.

[15] 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'l Conf. Multimedia*, 2014.

[16] 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'l Conf. Multimedia*, 2010.

[17] 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*, vol. 20, no. 1, pp. 112–134, 2015.

[18] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach, "Intersubject synchronization of cortical activity during natural vision," *Science*, vol. 303, no. 5664, pp. 1634–1640, 2004.

[19] 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. 22nd IEEE Int'l Conf. Image Processing (ICIP)*, 2015.

[20] A. Zlatintsi, E. Iosif, P. Maragos, and A. Potamianos, "Audio salient event detection and summarization using audio and text modalities," in *Proc. 23rd European Signal Processing Conference (EUSIPCO)*, 2015.

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

[22] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, 2005.

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

[24] D. A. Harville, "Maximum likelihood approaches to variance component estimation and to related problems," *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 320–338, 1977.

[25] K. J. Friston, A. P. Holmes, K. J. Worsley, J.P. Poline, C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human Brain Mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[26] W. D. Penny, A. P. Holmes, and K.J. Friston, "Random effects analysis," *Human Brain Function*, vol. 2, pp. 843–850, 2003.

[27] J. L. Mozolic, D. Joyner, C.E. Hugenschmidt, A. M. Peiffer, R. A. Kraft, J. A. Maldjian, and P. J. Laurienti, "Cross-modal deactivations during modality-specific selective attention," *BMC Neurology*, vol. 8, no. 1, pp. 35, 2008.

[28] A. C. Huk, R. F. Dougherty, and D. J. Heeger, "Retinotopy and functional subdivision of human areas MT and MST," *The Journal of Neuroscience*, vol. 22, no. 16, pp. 7195–7205, 2002.