Exploring CNN-based architectures for Multimodal Salient Event Detection in Videos

Petros Koutras, Athanasia Zlatinsi and Petros Maragos
School of E.C.E., National Technical University of Athens, 15773 Athens, Greece
Email: {pkoutras, nzlat, maragos}@cs.ntua.gr

***Abstract*—Nowadays, multimodal attention plays a significant role in many machine-based understanding applications, computer vision and robotic applications, such as action recognition or summarization. In this paper, we present our approach to the problem of audio-visual salient event detection based on visual and audio modalities by employing modern Convolutional Neural Network (CNN) based architectures. In this way, we extend our previous work, where a hand-crafted frontend was examined, an energy based synergistic approach, where a non-parametric classification technique was used for the classification of salient vs. non-salient events. Our comparative evaluations over the COGNIMUSE database [1], consisting of movies and travel documentaries, as well as ground-truth data denoting the perceptually mono- and multimodal salient events, provided strong evidence that the CNN-based approach for all modalities (i.e., audio, visual and audiovisual), even in this task, manages to outperform the hand-crafted frontend in almost all cases, accomplishing really good average results.**

I. Introduction

One of the most fundamental research challenges nowadays is the automatic video understanding that assists people with effective organization, retrieval, indexing, compression or even summarization of the video content. This has come to be eminent due to the increased amount of video data (i.e., movies, documentaries, home videos, music videos etc.) that have grown into an easily created and distributed media. People, in order to parse, structure and organize the available information, use cognitive mechanisms such as attentional selection and information abstraction that are grounded in conscious or non-conscious activities, such as guided search, communication and interaction, awareness, action taking, visual and auditory scene analysis etc. [2], [3].

Event detection for the task of video summarization and abstraction has been the subject of many recent research works aiming to create automatic summaries generated either by using key-frames, which correspond to the most important video frames [4]–[6], or with video skims that combine the most descriptive and informative video segments [7]–[9] (Fig. 1). To tackle the problem of summarization various algorithms have been proposed [5], [7], [8]. Some of these relate to user attention or saliency models [8], [10], they can be domain-specific [11], relate to the plot of the video [7], or the query context [12]. For more general reviews about video summarization we refer the reader to [8], [10], [13]–[15].

This research work was partially supported by the project "COGNIMUSE" which is implemented under the "ARISTEIA" Action.

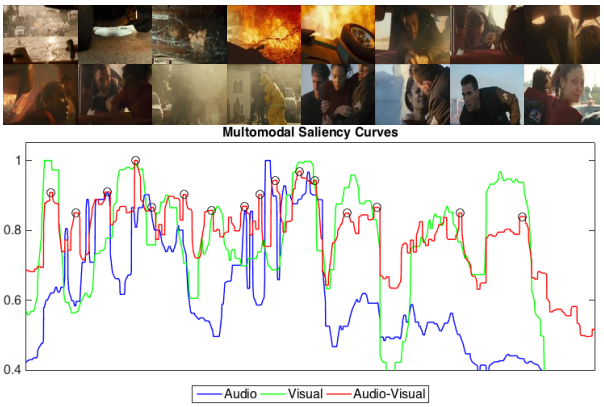

Fig. 1. Monomodal (A-blue, V-green) and Multimodal (AV-red) estimated Saliency Curves (bottom) and the respective Keyframes (top), extracted as local extrema of the AV curve for the movie Crash. Best viewed in color.

However, due to this increased number of video data, there is an immediate need to find video descriptors that solve large-scale video tasks. As stated in [16] such descriptors need to be generic, compact, efficient to compute and simple to implement. With today's breakthrough in the area of deep learning, which has been extensively applied in various image domains, giving state-of-the-art results on applications such as recognition [17], detection, segmentation [18] and retrieval, this can be now realized.

Various approaches [19]–[21] employ 3D-convolutional networks on short video clips of a few seconds in order to learn motion features from raw frames implicitly and then aggregate predictions at the video level. In [21] was shown that their network was marginally better than a single frame baseline. In [22] they incorporated motion information from optical flow, sampling up to 10 consecutive frames at inference time. In [23] a max-pooling Convolutional Neural Network (CNN) architecture as well as a recurrent neural network that uses Long Short-Term Memory (LSTM) cells are proposed to obtain global video-level descriptors for the combination of image information on full length videos and not just short clips, showing their ability to handle such videos.

The existing works on CNN architectures, which deal with video content can be classified into two main categories: 1) learning local spatiotemporal filters [16] and 2) incorporating optical flow using two-stream CNNs [22], [24]. In the first approach, the so-called C3D method learns a 3D CNN on a

978-1-5386-0951-4/18/$31.00

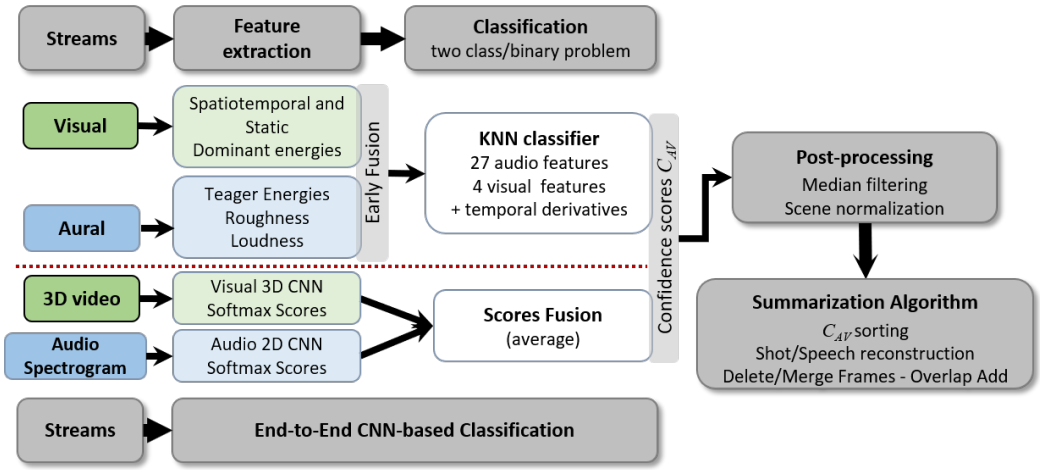

Fig. 2. System architectures overview for multimodal saliency detection and video summarization. The hand-crafted frontend (top-above the red dotted line) and the CNN-based arhitectures (bottom) can be seen.

limited temporal support of consecutive frames by letting all filters operate over space and time. On the other hand, two-stream CNN architectures decompose the video into spatial and temporal components by using RGB and optical flow frames. These components are fed into separate deep CNNs, to learn spatial as well as temporal information about the appearance and movement of a related visual activity. Each stream is performing the recognition on its own and for the final decision, softmax scores are combined by late fusion.

In this paper, motivated by the recent advances in deep architectures and the fact that multimodal framewise saliency (see Fig.1) can be another area where deep learning can be efficiently applied, we propose CNN-based architectures for salient event detection. For the visual stream we employ a CNN approach based on 3D convolutional nets (C3D), while for the audio stream a 2D CNN (based on the VGG idea of small kernels) is applied for this higher level salient vs. not salient event detection. Those architectures are compared to our baseline hand-crafted frontend for saliency detection and movie summarization, over our multimodal video database [1], consisting of mono- and multimodal ground-truth saliency annotations, which are crucial for training, content analysis and evaluation. The presented results show to be promising.

II. COGNIMUSE Multimodal video Database

The *COGNIMUSE Database* [1] is a video oriented database annotated with ground-truth annotations for sensory and semantic saliency, audio and visual events, cross-media relations as well as emotion, aiming to assist in training and evaluation of event detection and summarization algorithms. It consists of half-hour continuous segments from seven Hollywood movies[1] (three and a half hours in total), five travel documentaries[2], ca. twenty minutes long and a full length movie[3] with a duration of ca. hundred minutes. From the seven Hollywood movies, we excluded FNE as outlier, since it is an animation movie and thus not fitting for the training and evaluation of the CNN approach since we have not any other animation movies in the training set.

The ground-truth annotation of the database was based on video elements that captured the viewers' attention instantaneously or in segments including monomodal, i.e., audio (A) and visual (V) saliency annotation, and multimodal (AV) saliency annotation of the sensory content; hence, segments that are acoustically, visually or audio-visually interesting. In this paper, the sensory annotation was used for training and evaluation of our algorithms. For more details regarding the COGNIMUSE database we refer the reader to [1].

III. Architectures for Saliency Detection

Figure 2 shows the two system architectures for multimodal saliency detection and video summarization; the hand-crafted frontend (top) that employs state-of-the-art computational algorithms for feature extraction and salient event detection (top) and the End-to-End CNN-based architecture (bottom), combining softmax scores; both presented next.

A. Hand-crafted Frontend

For visual saliency estimation, we employed the recently proposed spatio-temporal model for visual saliency, which has achieved a good performance in many applications such as movie summarization [25], eye-fixation prediction [26] and action classification [27]. Specifically, we employed a spatio-temporal filterbank of 400 3D Gabor filters [26], as described in [25] for both the luminance and color streams to extract spatio-temporal and static dominant energies. This model is assumed to be more relevant to the cognition-inspired saliency methods [28], [29]. For the auditory modeling we employed the audio features proposed in [30]. These features are based on the Teager-Kaiser Energy Operator [31] and AM-FM demodulation [32], and variants have been successfully used in many applications such as speech and music recognition [32],

[1]List of movies: "A Beautiful Mind" (BMI), "Chicago" (CHI), "Crash" (CRA), "The Departed" (DEP), "Gladiator" (GLA), "Lord of the Rings - the Return of the King" (LOR) and the animation movie "Finding Nemo" (FNE).

[2]List of travel documentaries: four episodes from "Alternate Routes" series, i.e., "London" (LON), "Tokyo" (TOK), "Sydney" (SYD), "Rio" (RIO) and one episode from "Get Outta Town" series: "London" (GLN)

[3]Full movie: "Gone with the Wind" (GWW) (the first part)

Fig. 3. CNN architecture for visual saliency detection using deep 3D convolutional nets.

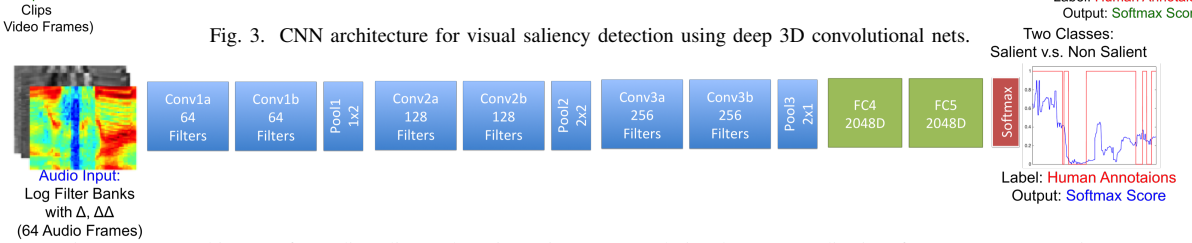

Fig. 4. CNN architecture for audio saliency detection using 2D convolutional nets on audio time-frequency representations.

[33] and summarization [8], [25], [30]. The employed features 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 and attention [34], are included as additional features. The combination of the visual and audio frontends constitute our energy based synergistic approach for audio-visual salient event detection and video summarization.

Machine Learning based architecture for the classification of salient vs. non-salient events: For the multimodal saliency event detection task a machine learning classification approach has been adopted, where we employed a K-Nearest Neighbor Classifier (KNN) as in [25]. The combination of the above mentioned features was used, thus, 4 visual and 27 auditory features, along with their first and second temporal derivatives. Specifically, we considered frame-wise saliency as a two-class classification problem, while a confidence score was also determined for every classification result (i.e., each frame), in order to obtain results for various compression rates and hence be able to produce summaries of various lengths. For the creation of the summaries [25], we have used the classifier's output that consists of the frames classified as salient; thus, segments or frames (chosen based on high confidence scores) are used as an indicator function curve, representing the most salient audio-visual events.

B. CNN-based architectures for saliency detection

Convolutional Neural Networks consist a biologically inspired class of deep learning models, which could actually replace the stages of feature extraction and classification to one single network that is trained end-to-end from raw pixel values to classifier outputs. Here, we propose two architectures for the estimation of the visual and audio saliencies as the softmax scores of the CNN output. We used the monomodal (A, V) and the multimodal (AV) annotations as ground-truth labels for the two classes (salient vs. non salient) during the training phases. For both networks we employ a Multinomial Logistic Loss for binary classification which takes the form:

$$\mathcal{L}(\mathbf{W}) = -\sum_{j \in Y_+} \log P(y_j = 1 | X; \mathbf{W}) - \sum_{j \in Y_-} \log P(y_j = 0 | X; \mathbf{W}),$$

where $\mathbf{W}$ are the trainable parameters of a CNN, X are the network input samples (see Figs. 3, 4), $y_j \in \{0, 1\}$ is the binary saliency label of X, and Y_+ and Y_- are the positive (salient) and negative (non-salient) labeled sample sets. $P(\cdot)$ is obtained by the softmax activation of the final layer. The trained models can then be employed for computing saliency curves in a new unseen video.

Visual 3D CNN: The core stage of the hand-crafted visual frontend consists of a spatio-temporal filtering with a carefully designed 3D Gabor filterbank. Then the extracted energies are sent to the classifier in order to take advantage of the existing ground-truth saliency annotations. In this work, we propose a deep end-to-end CNN architecture for learning the filterbank parameters as a sequence of 3D convolutional networks. We employed a CNN approach based on 3D convolutional nets (C3D) that was first introduced in [16], mainly for the action recognition problem. We believe that this approach is also suitable for the higher level concept of salient vs. non-salient events detection, since C3D nets can learn spatio-temporal patterns, which are related to visual saliency.

The 3D CNN has the ability to model successfully the temporal information of a video since convolutions and pooling operations are applied inside spatio-temporal cuboids, while in classic CNNs they are done only in the spatial domain. For this reason, the dimension of the feature maps in each convolutional layer is $n \times t \times h \times w$, where the additional parameter t stands for the number of video frames, while w and h describe the spatial size of each frame and n defines the number of filters in each layer. The employed deep C3D architecture is shown in Fig. 3. Specifically, videos are split into non-overlapping 16-frame RGB clips, which are used as input to the networks. The C3D network has 8 convolutional, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. The number of filters are denoted in each box. The sizes of all 3D convolutional kernels are $3 \times 3 \times 3$, and the stride of all these kernels are 1 in both spatial and temporal domain. The sizes of all pooling kernels are $2 \times 2 \times 2$, except for the first one. We used small kernel sizes following the idea proposed by VGG Net [17] to replace large convolutional kernels by a stack of 3×3 kernels without pooling between these layers. Each fully connected layer has 4096 output units and we have also applied dropout with a probability 0.5.

We trained from scratch end-to-end C3D models in Caffe [35] using an Nvidia Titan X GPU. We used mini-batches of

30 clips, with initial learning rate 0.003 and momentum 0.9. We have applied 10000 iterations of the Stochastic Gradient Descent, which approximately corresponds to 15 epochs. The learning rate is divided by 10 after the half of the iterations.

Audio 3D CNN: For the audio stream we employed a 2D CNN that is based on the VGG idea of small kernels and is also applied in other audio related problems, i.e., acoustic event detection [36]. In this approach we want to represent the raw audio signal in the 2D time-frequency domain and preserve locality in both axis. The conventional mel-frequency cepstral coefficients (MFCCs) representation cannot maintain locality to the frequency axis due to the discrete cosine transform projection. Thus, 50 log-energies were computed directly from the mfccs using 25 ms frames with 10 ms shift. In addition, we computed first and second temporal derivatives, in order to have a 3 channel 2D input, similarly to the RGB image.

Figure 4 shows the employed deep 2D CNN architecture. The network has 6 2D convolutional, 3 2D max-pooling, and 2 fully connected layers, followed by a softmax output layer. The number of filters are denoted in each box. The max-pooling layers are written as time $\times$ frequency. The sizes of all 2D convolutional kernels are 3×3, and the stride of all these kernels is fixed to 1. For the training we followed a similar to the visual approach; however, we used mini-batches of 128 clips, with initial learning rate 0.02 and momentum 0.9.

IV. Experimental Evaluation on COGNIMUSE DB

For the evaluation of the various video data included in the COGNIMUSE database, different types of evaluation setups were adopted. For the six Hollywood movies a six fold cross-validation was applied, where five movies were used for training and tested on the sixth. For the travel documentaries a five fold cross-validation was considered, where four documentaries where used for training and the fifth for testing, while for GWW two different setups were adopted; i) only the six Hollywood movies were used for training (GWW*) and ii) all data was used for training (GWW**), thus six movies and five travel documentaries.

Table I shows evaluation results using AUC (Area Under Curve) as a metric between the hand-crafted frontend (denoted as Hndcr.) and the CNN-based architectures for all modalities and evaluation setups, thus audio on audio (A-A), visual on visual (V-V) and audio-visual on audio-visual (AV-AV) annotation, for all movies (top) and travel documentaries (bottom) individually and on average for each video genre. For the audio-visual saliency estimation we fuse, using average, the softmax scores that are provided by the two-stream CNNs trained with the AV annotation labels.

For the movies, we immediately observe that the CNN-based architecture outperforms the hand-crafted frontend in almost all movies and all evaluation setups, e.g., for the visual modality we have an increase up to 3%. Note that for the audio modality only in CHI we cannot achieve an improvement due to the fact that this movie is a musical containing mostly music segments and thus the CNN training on the other movies cannot capture efficiently this type of information.

TABLE I
Evaluation results using AUC for the CNN-based architectures and the hand-crafted frontend (denoted as Hndcr.) comparing each modality to the corresponding annotation, i.e., V-V, A-A and AV-AV evaluation (using average for the CNN) individually for each movie and travel documentary. Average results are also shown. Regarding GWW, GWW* denotes results for training in the Hollywood movies only and GWW** results when the training was performed in all database videos.

AUC Results	**V-V**		**A-A**		**AV-AV (mean)**	
videos	**Hndcr.**	**CNN**	**Hndcr.**	**CNN**	**Hndcr.**	**CNN**
Six Hollywood Movies						
BMI	0.718	0.765	0.823	0.844	0.842	0.839
GLA	0.739	0.772	0.840	0.849	0.850	0.830
CHI	0.645	0.706	0.847	0.815	0.819	0.820
LOR	0.688	0.738	0.873	0.872	0.811	0.832
CRA	0.720	0.726	0.848	0.874	0.804	0.799
DEP	0.778	0.741	0.822	0.861	0.824	0.856
Aver.	0.715	**0.742**	0.842	**0.853**	0.825	**0.830**
Full Length Movie						
GWW*	0.589	0.644	0.714	0.706	0.664	**0.735**
GWW**	0.626	**0.660**	0.706	**0.740**	0.648	0.710
Five Travel Documentaries						
LON	0.650	0.806	0.794	0.830	0.777	0.814
RIO	0.668	0.718	0.690	0.737	0.821	0.805
SYD	0.621	0.771	0.726	0.787	0.734	0.863
TOK	0.767	0.831	0.796	0.849	0.819	0.856
GLN	0.657	0.679	0.809	0.894	0.693	0.810
Aver.	0.673	**0.761**	0.763	**0.819**	0.769	**0.830**

Significant improvement is also obtained for GWW, which constitutes a real challenging task since it is a full length movie. We note an improvement of up to 7% for the visual and audio-visual evaluation and more than 3% for the audio when trained in all videos. For the five travel documentaries, we note that the deep CNN-based architectures, for the visual modality, significantly outperforms the hand-crafted frontend with an increase up to 15% for LON and an average percentage of ca. 9% for all five travel videos, while an increased performance of ca. 6% is noted for audio and audio-visual modalities. Finally, we notice that even though the two networks are trained independently for the AV case, their late fusion of the softmax probabilities achieves a really good performance outperforming the baseline in most cases.

V. Conclusions

In this paper, we proposed CNN-based architectures for the problem of audio-visual salient event detection. These architectures were compared with our previous hand-crafted frontend, which employs advanced state-of-the-art methods for saliency detection, over the COGNIMUSE database, consisting of different types of videos and mono- and multimodal ground-truth annotations of the salient events. Our experimental evaluations show that the deep CNN-based architecture manages to outperform almost in all cases and all types of videos the hand-crafted frontend. For future work, we intent to further refine our methods and the proposed CNN architectures regarding their parameterizations by taking into consideration the transfer of the rapid CNN improvements that are achieved in other domains, i.e., visual object and action recognition.

REFERENCES

[1] 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*, vol. 2017, no. 1, p. 54, Aug 2017.

[2] M. I. Posner and S. E. Petersen, "The attention system of the human brain," *Annual Review of Neuroscience*, vol. 13, no. 1, p. 25?42, 1990.

[3] E. I. Knudsen, "Fundamental components of attention," *Annual Review of Neuroscience*, vol. 30, pp. 57–58, June 2007.

[4] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. Int'l. Conf. Computer Vision and Pattern Recognition*, 2012.

[5] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, "Large-scale video summarization using web-image priors," in *Proc. Int'l. Conf. Computer Vision and Pattern Recognition*, 2013.

[6] S. F. de Avila, A. B. Lopes, A. da Luz Jr., and A. de Albuquerque Araujo, "VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[7] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. Int'l. Conf. Computer Vision and Pattern Recognition*, 2013.

[8] 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. on Multimedia*, vol. 15(7), pp. 1553–1568, 2013.

[9] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 505–520.

[10] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. on Multimedia*, 2005.

[11] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, "Category-specific video summarization," in *Proc. European Conference on Computer Vision*, 2014. [Online]. Available: http://hal.inria.fr/hal-01022967

[12] M. Wang, R. Hong, G. Li, Z.-J. Zha, S. Yan, and T.-S. Chua, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Trans. on Multimedia*, vol. 14, no. 4, pp. 975–985, 2012.

[13] B. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, 2007.

[14] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Process. Mag.*, vol. 17, 2000.

[15] A. Money and H. Agius, "Video summarization: A conceptual framework and survey of the state of the art," *J. Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, Feb. 2008.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. International Conference on Computer Vision (ICCV)*, 2015.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int'l Conf. of Learning Representations (ICLR)*, 2015.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[19] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *2nd Int'l Workshop on Human Behavior Understanding (HBU)*, 2011.

[20] S. Ji, W. Xu, M. Yang, , and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. PAMI,*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[22] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2014.

[23] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[24] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: towards good practices for deep action recognition," in *Proc. European Conference on Computer Vision (ECCV)*, 2016.

[25] 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. Int'l Conf. on Image Process.*, Quebec, Canada, 2015.

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

[27] K. Maninis, P. Koutras, and P. Maragos, "Advances on action recognition in videos using and interest point detector based on multiband spatio-temporal energies," in *Proc. Int'l Conf. Image Processing*, 2014.

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

[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*, vol. 20(11), pp. 1254–1259, 1998.

[30] A. Zlatintsi, E. Iosif, P. Maragos, and A. Potamianos, "Audio salient event detection and summarization using audio and text modalities," in *Proc. European Signal Processing Conference (EUSIPCO-15)*, 2015.

[31] J. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Proc. IEEE Int'l. Conf. Acoust., Speech, Signal Process.*, 1990.

[32] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory teager energy cepstrum coefficients for robust speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, Lisbon, Portugal, Sep. 2005.

[33] A. Zlatintsi and P. Maragos, "AM-FM modulation features for music instrument signal analysis and recognition," in *Proc. 20th European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

[34] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*. Springer, 2nd edition, 1999.

[35] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *arXiv*, 2014.

[36] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," in *Proc. Interspeech*, 2016.