Improving Multimodal Movie Scene Segmentation Using Mixture of Acoustic Experts

Meng-Han Lin
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan
ea0270@gapp.nthu.edu.tw

Jeng-Lin Li
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan
cllee@gapp.nthu.edu.tw

Chi-Chun Lee
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan
cclee@ee.nthu.edu.tw

***Abstract*—Scenes are the most basic semantic units of a movie that are important as pre-processing for various multimedia computing technology. Previous scene segmentation studies have introduced constraints and alignment mechanisms to cluster low-level frames and shots based on the visual features and temporal properties. Recent researchers have extended by using multimodal semantic representations with the acoustic representations blindly extracted by a universal pretrained model. They tend to ignore the semantic meaning of audio and complex interaction between the audio and visual representations for scene segmentation. In this work, we introduce a mixture-of-audio-experts (MOAE) framework to integrate acoustic experts and multimodal experts for scene segmentation. The acoustic expert is learned to model different acoustic semantics, including speaker, environmental sounds, and other events. The MOAE optimizes the weights delicately among various multimodal experts and achieves a state-of-the-art 61.89% F1-score for scene segmentation. We visualize the expert weights in our framework to illustrate the complementary properties among diverse experts, leading to improvements for segmentation results.**

***Index Terms*—Movie, Scene Segmentation, Mixture of Experts, Multimodal Attention, Audio**

I. Introduction

Intelligent services, such as video retrieval, video understanding, and video tagging, aim at automatically handling rich multimedia data for entertainment services. Many of these services [1], [2] require reliable video segmentation to obtain semantically meaningful clips for the downstream tasks. Past studies have identified a lack of semantic consistency in the movie hierarchy using conventional units for movie segmentation, such as those based purely on frames or shots [2]. A basic semantic unit in a movie is defined as a scene, depicting a semantically cohesive segment of a story [3]. Scenes are tied to physical measures of narrative shifts and contain consecutive shots that form a high-level concept of events [4]. The boundaries typically coincide with the discontinuity of three factors: location, character, and time [4]. Due to the complexity of movie contexts related to these factors, a scene boundary cannot be easily identified with visual cues such as a cut. Tools for direct frames and shot detection, hence, are not applicable to the semantic scene segmentation task.

The importance of scene boundary detection is a key front-end processing for multimedia technology, various studies have developed algorithms using visual cues to segment scenes for movies. For example, Chen et al. have automated video editing rules for action scene segmentation [5]. Chasanis et al. have used low-level features to cluster similar shots into scenes [6]. Advanced algorithms have aligned sequences and imposed constraints to cluster temporally related shots [7]. Dynamic programming has further improved optimization of adjacent shot grouping by formulating a cost function of the global shot similarity matrix [8]. Another branch of algorithms regard the scene transition as a partition of a graph constructed by the visual similarity of shots [9]. To capture the semantics on the boundary, other studies have used pretrained models to derive high-level representations. For example, Baraldi et al. have considered places and objects semantics by extracting pretrained embeddings for a Deep Siamese Network [10]. Most of the previous researches focused on visual semantics while ignoring the acoustic aspect in movies.

Although some studies have computed low-level audio descriptors or pretrained embeddings [8], [11], [12] for scene segmentation, there was no sound multimodal approach until Rao et al. proposed a hierarchical framework using 4 types of multimodal pretrained representations. They established a state-of-the-art scene boundary identification framework using place, character, and action representations from visual contents, along with STFT of speech and background sound [13]. However, these acoustic features were derived from simple short-time fourier transformation which makes describing distinctively diverse semantics of audio non-intuitive. For instance, a representation modeling environmental sounds for the segmentation lacks harmonic properties of human speech or music, which are also crucial to determine scene changes. A change in speakers or music can suggest a transition point of character interactions or a story. A change in music signal a turning point of a story. That is, a framework that can integrate diverse visual and acoustic semantics delicately is crucial for scene segmentation of complex movie contents.

In this paper, we propose a mixture-of-acoustic-experts (MOAE) framework which integrates multiple representations to improve the multimodal scene segmentation. The MOAE framework contains acoustic experts, multimodal experts, and a mixture network, which generalizes from a multimodal framework [13] to enable fusion of multi-experts. The acoustic experts are learned by multi-aspect acoustic-based semantic

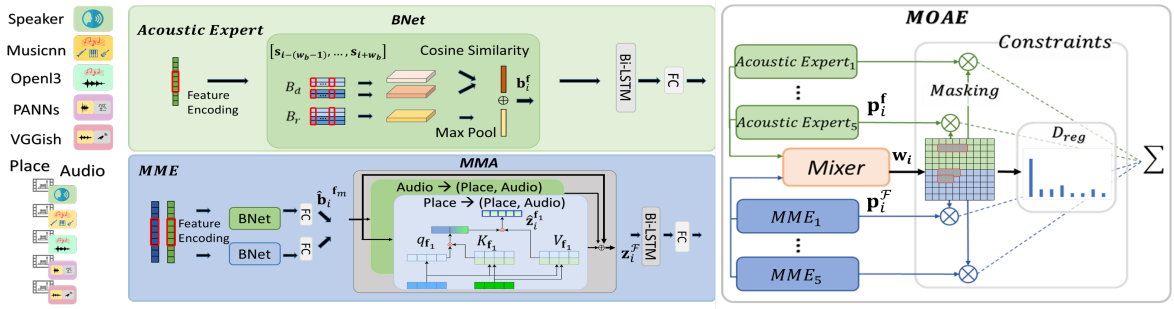

Fig. 1. The *acoustic expert*(section II-C1) can perform boundary prediction using various acoustic features. The *multimodal expert* (denoted as MME and described in section II-C2) combines *place* and acoustic features. We use five types of acoustic features for five *acoustic experts* and pair each of them with *place* to train five MME experts. The MOAE framework(section II-C3) deploys a mixer network with learning constraints using the prediction probabilities from five acoustic experts and five MME experts.

TABLE I
A SUMMARY OF EXTRACTED FEATURES.

Visual	**Pretrain Dataset**	**Aspect**
Place	Places365 dataset	Physical Scene
Cast	CIM Dataset, PIPA Dataset	Character instance
Action	AVA dataset	Character Action
Acoustic	**Pretrain Dataset**	**Aspect**
STFT	AVA-ActiveSpeaker	Time-localized Frequency
Speaker	LibriSpeech Corpus	Character Speech
VGGish	Youtube-100M Dataset	Acoustic Environment
PANNs	Audio Set	Acoustic Environment
Openl3	Music subset of Audio Set	Music Information
Musicnn	MagnaTagATune Dataset	Music Information

representations including speaker, music, audio events, and environmental sounds. A multimodal attention [14] module combines visual representations with acoustic representations. The mixture network provides further optimization of the fusion results. We evaluated the framework on a database including 1,110 scenes and obtained a 61.89% F1-score. We compare to the state-of-the-art multimodal scene segmentation work [13] and different ensemble learning approaches. Further analyses on mixer weights revealed the complementary property between acoustic and multimodal experts.

II. Database and Methodology

A. Database and Scene Annotation

Our database includes 10 movies ranging from 94 to 141 minutes. There are a total of 14k shots with 1,110 annotated scene boundaries. A movie scene a sequence of semantically cohesive shots. We follow a previous research [4] to form annotation guidelines for cohesive shot boundaries as follows. We first divide each movie into shots using a publicly available tool [15] and then label each shot by judging if its ending boundary is a scene boundary or not. Annotators identify a change of an event in a movie as a scene boundary by considering three factors including shifts of locations, characters, and time. Two annotators have labeled all the boundaries and achieved 95% label agreement.

B. Multimodal Shot Representations

We use a set of pretrained models to extract shot representations (summarized in Table I). Visual representations are extracted from key frames using pretrained models for place, cast, and action recognition tasks [13]. For acoustic semantics, we extract various acoustic embeddings using pretrained networks on different audio tasks. VGGish [16] is a pretrained convolutional neural network (CNN) for general audio classification. PANNs [17] is a pretrained Wavegram-Logmel-CNN on an audio tagging dataset, AudioSet [18].Both model embeddings have been applied to various tasks, such as environmental sound classification (ESC) and audio event detection tasks. We use Openl3 [19], [20], a self-supervised model, and Musicnn [21], [22], a music tagging model, to extract music-related embeddings. Finally, we use Deep speaker model [23] to extract speaker embeddings, which have been applied to speaker verification tasks.

A movie contains a sequence of N shots which are represented by extracted shot representations $[\mathbf{s}_1, ..., \mathbf{s}_i, ..., \mathbf{s}_N]$. We use an LSTM embedding layer to encode the temporal representations with its final hidden state $\mathbf{s}_i$. This layer is jointly optimized with the rest of the segmentation framework.

C. Framework

Our proposed MOAE framework comprises *acoustic experts*, *multimodal experts*, and a *mixer network* (figure 1). We compute a sequence of shots $\mathcal{S}$ of length L ($L << N$) as a boundary representation in the expert networks, and classify whether this boundary is a scene boundary. The mixture network is designed to combine decisions of the expert networks for an optimized prediction.

1) Acoustic Expert and Segment Prediction: An hierarchical local scene segmentation network (*HLSS*) [13] is used as a base network to identify semantic shifts on a boundary. Given a sequence of $2w_b$ ($w_b < L$) shots $[\mathbf{s}_{i-(w_b-1)}, ..., \mathbf{s}_{i+w_b}]$ for the boundary between the i-th and $i+1$-th shots, a shot-level boundary network (BNet) in *HLSS* embeds shot relations and differences for a boundary representation using a relation branch B_r and a difference branch B_d, respectively. B_r consists of a temporal convolution layer with max pooling to capture frame relations in a shot. B_d consists of two temporal convolution layers to learn representations of the current and the next shots, and outputs the cosine distance between the two representations. Each *HLSS* network using a type of acoustic features $\mathbf{f}$ is regarded as an *acoustic expert* which can generate

a concatenated embedding from B_r and B_d as a boundary representation $\mathbf{b}_i^{\mathbf{f}}$. The boundary representation is then passed through a Bi-LSTM layer followed by dense layers to generate a hidden embedding $E_{\mathbf{f}}$ and prediction by a softmax layer.

2) Multimodal Expert and Segment Prediction: We propose a *multimodal expert* to learn a audio-visual joint boundary representations using multimodal attention (*MMA*) [14]. With M types of feature, the feature set $\mathcal{F} = \{\mathbf{f}_1, ..., \mathbf{f}_m, ..., \mathbf{f}_M\}$ are encoded as d_{std} dimensional boundary representations $\hat{\mathcal{B}}_i^{\mathcal{F}} = \{\hat{\mathbf{b}}_i^{\mathbf{f}_1}, ..., \hat{\mathbf{b}}_i^{\mathbf{f}_m}, ..., \hat{\mathbf{b}}_i^{\mathbf{f}_M}\}$ using M distinct BNets. The cross feature attention representation is computed using $\hat{\mathcal{B}}_i^{\mathcal{F}}$ as input to learn self-attention and directional attention weight. For all $m \in M$, $\hat{\mathbf{b}}_i^{\mathbf{f}_m}$ is transformed into query vector $q_{\mathbf{f}_m}$ and is applied to key vector $K_{\mathbf{f}_m}$ and value vector $V_{\mathbf{f}_m}$, where $K_{\mathbf{f}_m}$ and $V_{\mathbf{f}_m}$ are the linear transformations of $\hat{\mathcal{B}}_i^{\mathcal{F}}$. $q_{\mathbf{f}_m}$, $K_{\mathbf{f}_m}$, and $V_{\mathbf{f}_m}$ are then used to compute cross feature representation $\hat{\mathbf{z}}_i^{\mathbf{f}_m}$: $\mathbf{f}_m \rightarrow (\mathbf{f}_1, ..., \mathbf{f}_m, ..., \mathbf{f}_M)$. Finally, $\hat{\mathbf{z}}_i^{\mathbf{f}_m}$, for all $m \in M$, are concatenated with skip connections to form $\mathbf{z}_i^{\mathcal{F}}$.

3) Mixer Network and Constraints: We train a mixer network to learn weights using experts' hidden representations for the fusion of the acoustic and multimodal experts. With N_E distinct expert models described in section II-C1 and II-C2, we use the hidden representations $E_{\mathbf{f}}$ of all N_E models as the input for the mixer network. The mixer network using a Bi-LSTM with dense layers which learns weights $\mathbf{w}_i \in \mathbb{R}^{N_E}$ to combine N_E distinct sets of prediction probabilities by weighted sum.

To mitigate the mixture of experts(MoE) overemphasizes the experts with higher accuracy [24], [25], we design three novel constraints to regularize the weight learning. The variance constraint D_{var} regularizes the expert weights $\mathbf{w}_i$ by maximizing the standard deviation of the weights. Scaled by the range R of the weights, D_{var} is expressed as follows:

$$D_{var} = 1 - \frac{1}{1 + tanh(\alpha * R * std(\mathbf{w}_i))}$$

The confidence constraint D_{conf} regularizes the probability difference between two classes, scene boundary $\mathbf{p}_{i_1} \in \mathbb{R}^{N_E}$ and non-scene boundary $\mathbf{p}_{i_0} \in \mathbb{R}^{N_E}$. D_{conf} penalizes too much weight assignment to the expert predicting high prediction confidence.

$$D_{conf} = 1 - \frac{1}{1 + tanh(\alpha * R_{md} * std(\mathbf{w}_i \odot |\mathbf{p}_{i_0} - \mathbf{p}_{i_1}| * \beta))}$$

$$R_{md} = max(\mathbf{w}_i \odot |\mathbf{p}_{i_0} - \mathbf{p}_{i_1}| * \beta) - min(\mathbf{w}_i \odot |\mathbf{p}_{i_0} - \mathbf{p}_{i_1}| * \beta)$$

α and β are constants used to scale the loss terms. Each of the constraints can be integrated into a new loss function through:

$$Total\ Loss = Loss_{ce} * (1 + D_{reg}),$$

where D_{reg} can be D_{var} or D_{conf}, and $Loss_{ce}$ is the cross entropy loss. Aside from the constraints, we introduce an additional masking mechanism (denoted as $Masking$) as a constraint on the prediction values of the experts. The mask is designed to regularize too many scene boundaries predicted in the shot sequence. Specifically, we aggregate positive class prediction probabilities of consecutive scene boundaries as a consecutive score for each expert. We mask out an expert with a zero weight if the expert has top *K consecutive scores*.

III. Experiment

A. Experimental Setup

We conduct 5-fold cross validation experiments and evaluate classification results using average precision (AP), intersection of union (Miou), recall, precision, and F1-score.

- Exp. I: Comparison of unimodal and multimodal experts
- Exp. II: Comparison of ensemble approaches
- Weight analysis: Examining the weights from the mixer network in *Exp II* and the corresponding movie shots.

In Exp. I, $HLSS_X$(section II-C1) and MME_X(section II-C2) denote expert models using different features, where X indicates the type of feature described in section II-B or the combinations of multiple types of features. $HLSS_{combine}$ includes $place, cast, act, STFT$ features and is regarded as the most current state-of-the-art baseline for the multimodal scene segmentation task [13].

In Exp. II, we use the mixer network including 10 experts marked with "*" in table II to compare various constraints and ensemble approaches as follows:

- $Majority\ Voting$: Majority voting using expert predictions from the 10 expert models.
- $Unweighted\ Averaging$: The unweighted average of prediction probabilities of 10 expert models.
- $Stacking$: Inputting ten expert models' prediction probability to support vector machine (SVC) or logistic regression (LR) as a meta-learner for boundary classification
- $MOAE_Y$: Mixture of acoustic experts using constraint Y described in section II-C3

We train the $HLSS$ and MME frameworks for 60 epochs with Adam optimizer and a weight decay of 0.0005. The learning rate is chosen from $\{10^{-3}, 10^{-4}\}$ and is divided by 10 at the 30^{th} epoch. The mixer network is trained for 20 epochs with Adam optimizer and a weight decay of 0.005. The learning rate is 10^{-4}. The weight ratio on losses for non-scene and scene boundary is 1:10. L and w_b are 10 and 2. α is chosen from $\{0.1, 1, 10\}$, β is 10 and *K* is 3.

IV. Results

A. Comparison of Unimodal and Multimodal Experts

We first compare to the state-of-the-art approach [13] using their proposed single modality and multimodal features, and demonstrate the results in the first part of Table II. The results indicate that scenes are highly related to the physical space, i.e., $HLSS_{place}$ outperforming the same framework trained with three other features extracted from our movie database. Interestingly, $HLSS_{combine}$ and $HLSS_{place}$ performed similarly even though $HLSS_{combine}$ used four semantic features (3 visual + 1 audio) as shot representations. That is, simple multimodal fusion cannot effectively leverage the complementary information among the features. The middle part of table II are the results of the $acoustic\ experts$. $VGGish$, $PANNs$, and $Openl3$ provide comprehensive details describing general

TABLE II
THE RESULTS OF EXP I, INCLUDING $HLSS$(SECTION II-C1) AND MME(SECTION II-C2).

Method	AP	Miou	Recall	Precision	F1
$HLSS_{place}$	53.7	52.3	74.93	35.36	47.48
$HLSS_{cast}$	31.9	44.1	64.29	24.81	34.97
$HLSS_{action}$	30.3	43.5	64.73	22.69	33.14
$HLSS_{STFT}$	13.3	13.7	6.68	8.56	6.46
$HLSS_{combine}$	51.9	52.5	74.45	35.07	47.27
$*HLSS_{PANNs}$	44.0	47.0	73.57	26.87	39.19
$*HLSS_{speaker}$	24.8	41.2	56.36	19.95	29.04
$*HLSS_{Musicnn}$	27.1	40.7	53.69	22.56	31.26
$*HLSS_{Openl3}$	41.3	49.9	59.21	31.65	40.73
$*HLSS_{VGGish}$	45.9	46.5	75.91	26.77	39.11
$*MME_{place\_speaker}$	61.1	54.3	78.71	36.69	49.74
$*MME_{place\_PANNs}$	66.2	57.7	80.35	40.22	53.08
$*MME_{place\_Musicnn}$	56.6	54.7	75.3	37.9	50.00
$*MME_{place\_Openl3}$	63.9	58.2	75.95	44.46	55.45
$*MME_{place\_VGGish}$	61.8	59.2	70.28	45.87	54.05

Methods marked with "*" are used to extract hidden representations and prediction probabilities in Exp. II.

environment and events of a movie scene from audio tracks. These three experts achieved higher performance than cast and action experts. In contrast to the general-purpose acoustic representations, $Musicnn$ and $speaker$ provides the information for more specifically designed scene transition techniques. The lower part of table II shows the results of MME models. MME_{place_PANNs} obtained 27.6% and 23.3% relative AP improvement over $HLSS_{combine}$ and $HLSS_{place}$, and MME_{place_Openl3} achieved 17.3% and 16.8% relative F1 improvement, respectively. $MME_{place\_speaker}$ and $MME_{place\_Musicnn}$ obtained 13.8% and 5.4% relative increased AP compared to $HLSS_{place}$, and improved F1 relatively by 4.8% and 5.3%, respectively. In contrast to $HLSS$ models, MME enables the model to integrate semantic information from various modalities. For example, the multimodal $HLSS$ degrade the performance using $Musicnn$ and $speaker$ due to dominant performance of $place$ while MME has learnable attention weights to flexibly address this issue.

B. Comparison of Ensemble Approaches

Table III shows the overall results of *Exp. II*. $MOAE$, the mixer network without constraints, leveraged the multimodal expert diversity and improved segmentation performance with 57.36% F1-score. Both $MOAE_{D_{var}}$ and $MOAE_{D_{conf}}$, designed to limit the imbalance of mixer weights, introduced 3.7% relative improvement over $MOAE$ in F1-score. While $MOAE_{Masking}$ improved F1 and precision score, recall and AP slightly decreased. This indicates the fact that correct predictions could sometimes be masked due to consecutive scene boundaries. $MOAE_A$, jointly constrained by D_{var} and $Masking$, further increases 7.9% relative F1-score compared to $MOAE$. $MOAE_B$ applying both D_{conf} and $Masking$ to the mixer network can also attain improvements. The decision fusion methods ($Majority\ Voting$ and $Unweighted\ Averaging$) outperform either $MOAE_{D_{var}}$ or $MOAE_{Masking}$. However, when both constraints simultaneously applied, $MOAE_A$ surpasses the best decision fusion method, with 2.9% relative improvement in F1-score. MOAE provides an edge over other multimodal fusion methods.

TABLE III
THE RESULTS OF EXP II. A DENOTES CONSTRAINTS WITH D_{var} AND $Masking$, AND B DENOTES CONSTRAINTS WITH D_{conf} AND $Masking$.

Method	AP	Miou	Recall	Precision	F1
$MOAE$ (section II-C3)	65.2	59.5	76.72	45.56	57.36
$MOAE_{D_{var}}$	66.6	61.2	77.62	48.79	59.47
$MOAE_{D_{conf}}$	66.9	61.4	77.29	49.02	59.47
$MOAE_{Masking}$	63.0	60.4	73.65	48.68	58.05
$MOAE_A$	66.2	63.3	74.48	53.70	61.89
$MOAE_B$	66.0	62.7	73.65	52.94	61.14
$Majority\ Voting$	-	61.5	73.92	50.59	59.66
$Unweighted\ Averaging$	65.7	56.9	77.46	49.71	60.16
$Stacking\ LR$	64.0	61.7	75.20	50.19	59.82
$Stacking\ SVC$	-	61.7	76.41	49.95	60.06

C. Weight Analysis

We plot weights of $MOAE_A$ along with three corresponding movie shot sequences in figure 2. All three shot sequences include a scene boundary denoted by a yellow line. The first two shots of the first row are the end of a conversation scene. By the end of the speech, the latter scene then begins a narration with background music. The sequence of shots in the narration alters frequently between various shooting location. Under this circumstance, the acoustic transition points apparently support the decision of scene boundary where the visual content is likely to create a false boundary. $MOAE_A$ assigned more weight to the acoustic experts to identify a scene boundary between these scenes. A similar weight distribution is observed at the next boundary, where the first and the second shot of the narration scene are in different locations. The second row depicts a different situation. The first two shots of the latter scene is used to establish the location of the scene, showing the exterior of a boat near the dock. These two shots are visually different from the following shots that show the interior of the boat. More weight is assigned to the acoustic experts because of the obvious visual variations. The third row is an example of visual variation caused by a drastic change in shot angle and shot type. This variation forced $MOAE_A$ to assign over 50% of total weight to acoustic experts even though these shots are all filmed in the same room.

V. CONCLUSION

In this work, we computed various acoustic semantic representations to complement the visual content modeling to improve the scene segmentation task. Moreover, we extended the multimodal hierarchical scene segmentation framework using the mixture-of-experts approach. The comparison experiments demonstrated state-of-the-art performances using our proposed MOAE framework. The expert weight analyses also visualized the impact of different aspects of the modalities on the decision of a boundary. To the best of our knowledge, this is one of the first works that comprehensively integrate information from

Fig. 2. Mixer weights and three corresponding shot sequences in the movie. A yellow line represents a scene boundary and the weights assigned to the experts for each boundary are shown between shots. These three sequences illustrate how the mixer assigns more weight to the acoustic experts when visual attributes vary dramatically, causing confusion when we rely on visual features.

audio tracks that also demonstrate a significant improvement beyond predominantly visual-based approaches in the task of scene segmentation for movies. We hope that this study can enable high-quality scene segmentation by including more semantics details, and therefore accelerate various machine learning applications in the multimedia domain.

REFERENCES

[1] Bhavesh Patel and B Meshram, "Content based video retrieval systems," *International Journal of UbiComp*, vol. 3, 05 2012.

[2] Aasif Ansari and Muzammil H Mohammed, "Content based video retrieval systems-methods, techniques, trends and challenges," *International Journal of Computer Applications*, vol. 112, no. 7, 2015.

[3] Ephraim Katz, *Ephraim Katz's The Film Encyclopedia*, Thomas Y. Crowell, 1979.

[4] James E Cutting, "Event segmentation and seven types of narrative discontinuity in popular movies," *Acta psychologica*, vol. 149, pp. 69–77, 2014.

[5] Lei Chen and M Tamer Ozsu, "Rule-based scene extraction from video," in *Proceedings. International Conference on Image Processing*. IEEE, 2002, vol. 2, pp. II–II.

[6] Vasileios T Chasanis, Aristidis C Likas, and Nikolaos P Galatsanos, "Scene detection in videos using shot clustering and sequence alignment," *IEEE transactions on multimedia*, vol. 11, no. 1, pp. 89–100, 2008.

[7] Rameswar Panda, Sanjay K Kuanar, and Ananda S Chowdhury, "Nyström approximated temporally constrained multisimilarity spectral clustering approach for movie scene detection," *IEEE transactions on cybernetics*, vol. 48, no. 3, pp. 836–847, 2017.

[8] Daniel Rotman, Dror Porat, Gal Ashour, and Udi Barzelay, "Optimally grouped deep features using normalized cost for video scene detection," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 187–195.

[9] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso, "Temporal video segmentation to scenes using high-level audiovisual features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 8, pp. 1163–1177, 2011.

[10] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara, "A deep siamese network for scene detection in broadcast videos," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1199–1202.

[11] Naoki Nitanda, Miki Haseyama, and Hideo Kitajima, "Audio signal segmentation and classification for scene-cut detection," in *2005 IEEE International Symposium on Circuits and Systems*. IEEE, 2005, pp. 4030–4033.

[12] Seungmin Rho and Eenjun Hwang, "Video scene determination using audiovisual data analysis," in *24th International Conference on Distributed Computing Systems Workshops, 2004. Proceedings.* IEEE, 2004, pp. 124–129.

[13] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin, "A local-to-global approach to multi-modal movie scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10146–10155.

[14] Zexu Pan, Zhaojie Luo, Jichen Yang, and Haizhou Li, "Multi-Modal Attention for Speech Emotion Recognition," in *Proc. Interspeech 2020*, 2020, pp. 364–368.

[15] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin, "Movienet: A holistic dataset for movie understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[16] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[17] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[18] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[19] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[20] Relja Arandjelovic and Andrew Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[21] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik M. Schmidt, Andreas F. Ehmann, and Xavier Serra, "End-to-end learning for music audio tagging at scale," in *19th International Society for Music Information Retrieval Conference (ISMIR2018)*, 2018.

[22] Jordi Pons and Xavier Serra, "musicnn: pre-trained convolutional neural networks for music audio tagging," in *Late-breaking/demo session in 20th International Society for Music Information Retrieval Conference (LBD-ISMIR2019)*, 2019.

[23] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[24] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017, OpenReview.net.

[25] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever, "Learning factored representations in a deep mixture of experts," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2014.