Level fusion analysis of recurrent audio and video neural network for violence detection in railway

Tony Marteau[*†], David Sodoyer[†], Sebastien Ambellouis[†], Sitou Afanou[*]
[*] SNCF Voyageurs, Centre d'Ingénierie du Matériel, Le Mans, France
{tony.marteau, sitou.afanou}@sncf.fr
[†] Univ Gustave Eiffel, COSYS-LEOST, F-59650 Villeneuve d'Ascq, France
{tony.marteau, david.sodoyer, sebastien.ambellouis}@univ-eiffel.fr

***Abstract*—This paper deals with the security improvement of passengers in public transport by automatically processing the audio and video streams of an embedded surveillance system. In this paper we analyse several levels of fusion of two deep audio and video recurrent network models for violent actions recognition. Each audio and video model is based on recent generic feature extractors proposed in the state-of-the-art to benefit of powerful feature representation capabilities. Each level of fusion is trained and evaluated on a new real-world audio-video surveillance streams recorded in a real train with scenes of violence played by actors. The obtained results confirm the interest in seeking to detect violence by jointly using audio and video signal and highlight the difficulty to define the optimal level of fusion.**

***Index Terms*—violence detection, audio-visual fusion, deep learning**

I. Introduction and Related Work

This paper deals with the automatic recognition of rare violent actions in a transport environment using multi-modal data. The automatic recognition of violent actions has been addressed for several years by modelling video streams using machine learning [1]–[3] and more recently using deep learning [4]–[8]. Moreover, automatic sound scenes and events recognition is being actively investigated [9] and several research deals with recognition and detection of violent scenes and screams [10]–[12].

In the specific context of transport safety and security, some studies have been proposed using either video stream [13], or audio stream [14]–[17], and either both stream but with independent models [18], [19].

At the same time, research is moving towards the processing of multi-modal signals [20] and fusion approaches have been experimented in use cases like speech enhancement [21], emotion recognition [22], tracking of multiple speakers [23], action recognition [24] or scene classification [25]. Because no open synchronized audio-video dataset for transport applications exists, these techniques have not been experimented for violence detection in this environment. In a transport environment, the processing of multi-modal signals for the detection of violence is a challenging problem because of occultations by the arrangement or by the passengers and also because of violence that occurs off-camera. To evaluate the interest in multi-modal signal processing, we propose a system for the recognition of violent actions in a transport environment with an audio-visual deep neural networks. For this, we also present a new audio-visual database composed of violent scenes recorded in a real railway environment.

II. Database

Few public databases containing scenes of violence in a railway environment exist in the literature and do not contain much data. We have therefore chosen to record our database in a double-deck suburban train operated by SNCF for the Ile-de-France region on the Transilien N line. Multiple violence scenarios were played by 18 professional actors between 9 am and 5:10 pm. Moreover, to have the transfer of the image rights of all the people in the train, no regular passenger was allowed in the train. The regular passengers were therefore played by 17 SNCF employees. For data recording, we have chosen not to use the installed CCTV system to have better quality data. Two mobile cameras with microphones (AXIS P3935-LR) are therefore placed longitudinally at each end of the rooms. The cameras are placed to reach the operational field of view of the CCTV cameras installed. Audio and video streams are acquired synchronously and respectively in 32bits at 44.1kHz for audio and 1280x720 at 25fps for video. Three different areas of the train are equipped with our recording equipment, illustration on the Figure 1. Two first areas are the lower and the upper passengers room of the train. The last area is the exchange platform with passenger room of the simple-deck part of the train. Each violence scenario is acted with different duration from 1 to 5 minutes in each room by two or more actors at three distances from the sensor: close-distance (0m to 3m), middle-distance (3m to 6m), and far-distance (6m to 9m). Moreover, each scenario was acted with different densities of regular passengers, from low density to heavy density. In total, 101 violent scenarios acted has allowed us to acquire 202 recordings with violence thanks to the cameras crossed field of view. Moreover, 11 scenarios without violence acted have allowed us to acquire 22 recordings thanks to the cameras crossed field of view.

III. Models

A. Features extraction

As our database is not large enough to train an end-to-end neural network, we have chosen to use generic feature extractors pre-trained by a large amount of data.

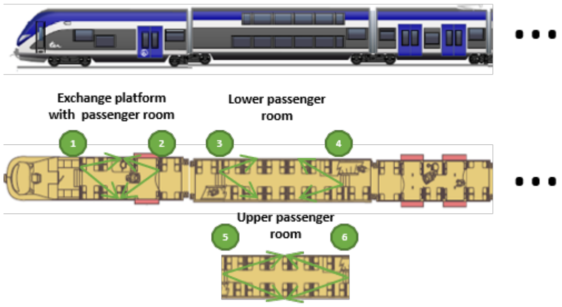

Fig. 1: Plan of a Regio2N

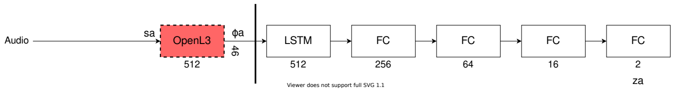

Fig. 2: Audio model

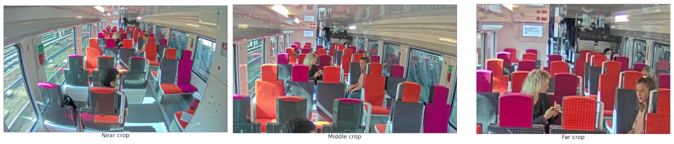

Fig. 3: Visualisation of cropping

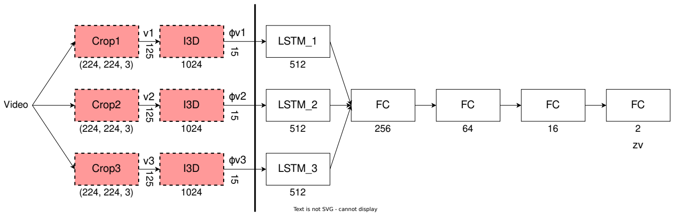

Fig. 4: Video model

For audio signal, we selected the audio OpenL3 model [26], [27]. This features extractor trained on AudioSet-Instruments database [28] is widely used in audio and audio-visual fusion communities for acoustic scenes or event classification [25], audio-visual correspondence task [27], or audio-visual scene classification [29].

For video signal, we did not select the visual OpenL3 model because it was not trained for action recognition. We therefore preferred to select RGB branch of I3D [30] pre-trained on Kinetics 400 database [31] designed for human action recognition.

B. Audio baseline

The audio baseline architecture is shown in Figure 2. We have designed a simple model which takes as input the features extracted by the selected OpenL3 model. These features are fed to a normalization batch layer, to obtain homogeneous features with a 0-centered Gaussian distribution, and then traited by a 512 units LSTM. The output of the LSTM is then provided to three fully connected layers activated by a ReLu, with respectively 256, 64, and 16 units. Finally, the output of the 16 units fully connected layer is passed into a two-unit classification layer activated by a softmax (cf. Equation 1), one unit is activated when there is no violence and the other unit is activated when there is violence in the input instance. In addition, a dropout layer is used between all the previously mentioned layers.

$$p_{audio} = softmax(z_a) \tag{1}$$

C. Video baseline

The video baseline architecture is shown in Figure 4. The proposed architecture takes into account the distance to the camera at which the scene takes place because of our railway environment: one input branch has been integrated for each distance. As previously specified, we use the I3D feature extractor as input of the video model. The extractor is applied to three different cropping areas, Figure 3. The first cropping is corresponding to the entire image for the close distance action, the second cropping centered on the center of the room, and finally the last cropping centered on the bottom of the room. The three cropping areas provided as input to the video baseline model are processed as three separate branches. Each branch is composed of one normalization batch layer and one LSTMs with 512 units. The outputs of the three LSTMs are merged by concatenation and then provided to three fully connected layers activated by a ReLu, with respectively 256, 64, and 16 units. Finally, as for the audio model, the output of the 16 units fully connected layer is passed into a two-unit classification layer activated by a softmax (cf. Equation 2). In addition, a dropout layer is used between all the previously mentioned layers.

$$p_{video} = softmax(z_v) \tag{2}$$

D. Fusion

Both audio and video output are fused as shown in Figures 5, 6, 7. We propose three level of fusion. The first fusion level (cf. Figure 5) consists of a simple fusion of the predictions from the uni-modal baseline models as described by Equation 3. In this case, if an alert is raised by a uni-modal model it will be raised by the predictions fusion system. The second fusion level is a late fusion on the penultimate fully connected layer (cf. Figure 6). And finally, the third fusion level is an early fusion after the uni-modal LSTM layers (cf. Figure 7).

$$p_{pred\_fusion} = \begin{cases} 0, & \text{if } p_{audio} = 0 \text{ and } p_{video} = 0 \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

On the first layers, the early and late fusion model processes the input features separately and in parallel like the video baseline model, i.e. by four layers (one for audio and three

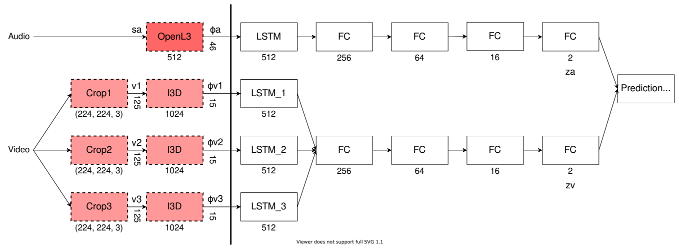

Fig. 5: Prediction fusion model

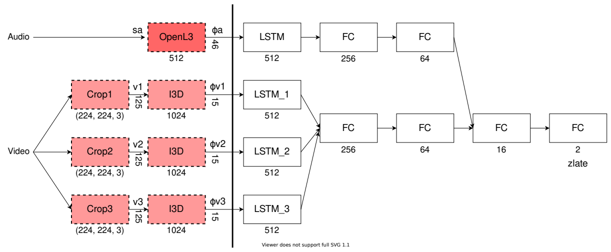

Fig. 6: Late fusion model

for video), with batch normalization and by four 512 units LSTM layers. Our fusion models vary after these layers.

For the late fusion model (cf. Figure 6 & Equation 4) separate and in parallel two fully connected layers are added after the audio LSTM layers, and two fully connected layers are added after the merged by concatenation of video LSTM. These fully connected layers have 256 and 64 units activated by ReLu. Concatenation fusion of signals is performed with the outputs of the 64 units fully connected layers. The result of the merge is then passes into a 16-unit fully connected layer activated by a ReLu and then passes into a two-unit classification layer activated by a softmax.

$$p_{late\_fusion} = softmax(z_{late}) \tag{4}$$

For the early fusion model (cf. Figure 7 & Equation 5) concatenation fusion is performed with the outputs of the LSTM layers. The result of the fusion passes then in three fully connected layers activated by a ReLu, with respectively 256, 64, and 16 units. Finally, as for the baseline models, the output of the 16 units fully connected layer is passed to a two-unit classification layer activated by a softmax.

In addition, a dropout layer is used in all fusion models between all the previously mentioned layers.

$$p_{early\_fusion} = softmax(z_{early}) \tag{5}$$

IV. Experience

A. Data

Our database has been annotated on the uni-modal perception of violence by following the procedure below. First of all, all our recordings have been cut into 2s segments, each segment was listened and viewed separately to get a violence annotation on each mode. This choice was made because the violence is not necessarily perceptible simultaneously on both modes. This mode's inconsistency can be explained for several reasons, violence without shouting or violence off-camera for example. Annotations of each mode are joined by following the Equation 6, with 0 for normal instance and 1 for violence instance, to form an event annotation. These event annotations have been used for the training and evaluation of models.

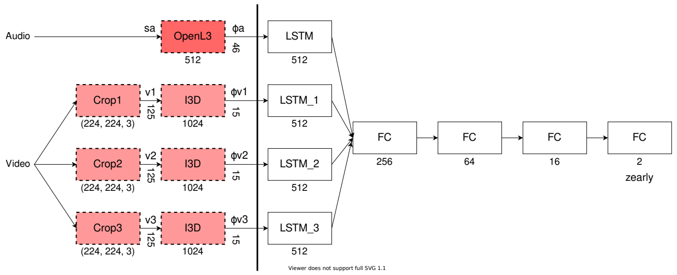

Fig. 7: Early fusion model

	Train set	Val set	Test set
Normal scenes	1574	202	365
Violence scenes	1782	284	411

TABLE I: Instances distribution in sets

$$y = \begin{cases} 0, & \text{if } y_{audio} = 0 \text{ and } y_{video} = 0 \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

Among the 224 recorded scenarios, 202 contain violence and 22 do not contain violence. These scenarios were cut into 5s instances with a hop length of 2s. We obtain 16939 instances with a distribution of 64% of normal instances and 36% of violence instances. This imbalance is explained by the fact that the violent phases represent only a small part of the scenarios played. To remove a training bias, we reduced this imbalance by applying random under-sampling on the normal class. The final instances distribution according to the sets is detailed in Table I.

Then features were extracted by the selected feature extractors. For the audio, we have provided 220500 samples representing 5s as input to the OpenL3 network. The extracted features are in $(46, 512)$ shape for the time-step and channels. For each crop of the video, we provide as input to the I3D network, selected for feature extraction, 125 RGB frames in $16:9$ format resized to 224×224. For video, feature extraction result in a $(15, 1024)$ shape output.

B. Training and evaluation

All classification models are trained with a batch size of 32 for 500 epochs with 0.33 dropout by an Adam optimizer with a learning rate of 0.0001 according to the MultiLabelSoftMarginLoss function:

$$loss(p, y) = -\frac{1}{C} * \sum_i y[i] * \log((1 + \exp(-p[i]))^{-1}) + (1 - y[i]) * \log(\frac{\exp(-x[i])}{(1 + \exp(-x[i]))}) \tag{7}$$

where p denotes the model prediction and y denotes the ground truth of input instance. The feature extraction part of the architectures (Figure 2,4,5,6,7) are not trained with our dataset.

Few works dealing with audio-visual data to detect events is available in the community to compare our models performance. We have therefore only select the AVE architecture, proposed by Tian et al. in [32]. This architecture has been proposed by authors for audio/video event localization in unconstrained videos. We have trained it on our database and compared it with the fusion models previously described.

Both uni-modal models are trained and evaluated to quantify how a fusion step is improving the global performance.

All models are trained by using the same parameters and the following procedure. As shown in Table I, the train set is used to learn the models weights to our task. The validation set is used to retain the best models weights with the minimum loss during the training. And finally, the evaluation is performed on the test set that is not seen during the training phase.

To evaluate the performance of all the models we use the global accuracy rate defined by:

$$accuracy = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{8}$$

where TP, FP, TN and FN are respectively the number of true positive, false positive, true negative and false negative predictions. This metric is relevant because we also balanced the test set.

In addition we use the confusion matrix and the Venn diagram [33] to describe more precisely the behaviour of our models w.r.t. the violence and no violence classes.

V. Performances

The Table II is regrouping the global accuracy rate for each architecture. First, by analyzing the accuracy of the uni-modal models, we can notice that the detection of the violence event is 8% more efficient by processing the audio signal than by processing the video signal. These better performances are explained by two factors. The first one is that the violent scenes are largely different from the normal scenes on the audio signal (quiet room vs. shouting). The second reason is that the video signal is sensitive to many occultations. For the predictions based fusion model, we notice that the performances are identical or less good than the uni-modal system: raising all the alerts of uni-modal systems seems not to be the optimal decision rule. The late and the early fusion models improve the accuracy w.r.t. the audio uni-modal model. The early fusion is 6% better than late fusion and early fusion model is 2% better than AVE architecture. Moreover, by analyzing the confusion matrices (cf. Figure 8), we observe that the late fusion model is better than early fusion model at reducing FP while the early fusion model is better at reducing FN.

By observing the Venn diagrams of these models, we try to go further in the analysis of errors. Each Figure 9 and 10, is representing the common errors (FP and FN) between uni-modal models and respectively early and late fusion model. Firstly, we can observe in the yellow circle of Venn diagram that early fusion makes only 12 new errors while late fusion model make 34 new errors. These new errors being less numerous is a major advantage for this architecture. Moreover, we can observe through the overlaps that the late model makes many common FN errors (74+22) with the video model, unlike the early fusion model (20+23) has common errors which proves that early fusion allows a better modelling between audio and video.

Model	Accuracy
AVE	86%
Video	73%
Audio	81%
Uni-modal prediction fusion	77%
Late fusion	82%
Early fusion	**88%**

TABLE II: Accuracy performance of models

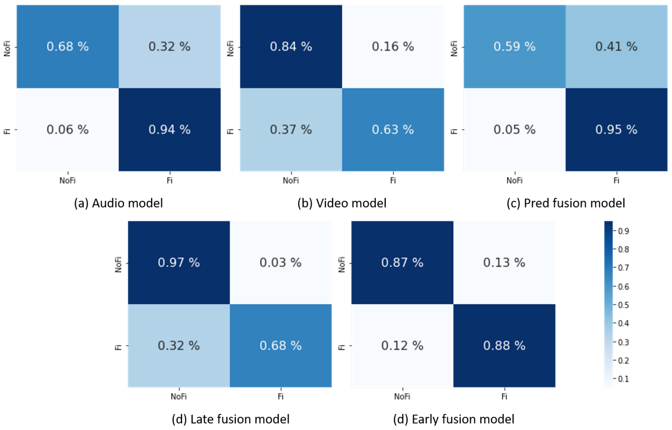

Fig. 8: Confusion matrix of models

VI. Conclusions

In this paper, we present a new railway audiovisual database containing violent scenes acted and acquired in a real train. We propose several audio/video LSTM based architectures based on two generic audio and video features extractor trained on large communities database. We evaluate the impact of the fusion level on the performance of the network that we compare with AVE architecture, a well-known state-of-the art

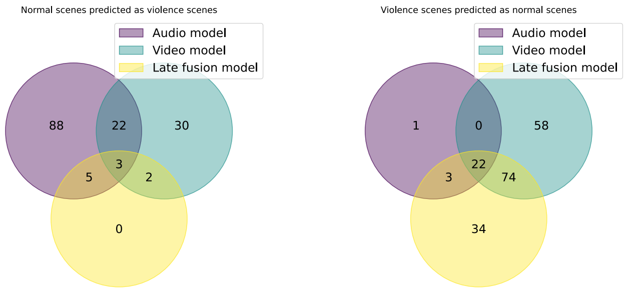

Fig. 9: Late fusion errors venn diagram

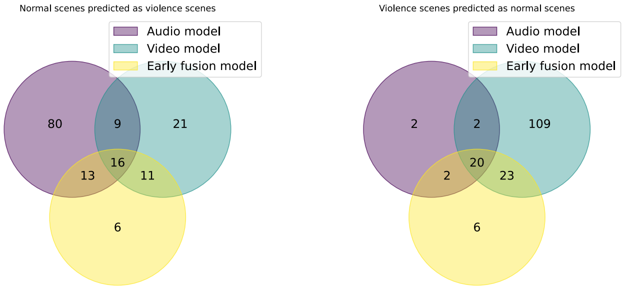

Fig. 10: Early fusion errors venn diagram

method. First, we show that without transfer learning of the feature extraction backbone, our architecture yield quite good results. Moreover, our evaluation shows that our early fusion architecture provide the highest global accuracy and better accuracy than AVE architecture.

References

[1] A. Datta, M. Shah, and N. da Vitoria Lobo, "Person-on-person violence detection in video data," in *Int. Conf. on Pattern Recognition*, Quebec, Canada, Aug 2002, pp. 433–438.

[2] E. B. Nievas, O. Déniz-Suárez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *Int. Conf. Computer Analysis of Images and Patterns*, Seville, Spain, Aug 2011, pp. 332–339.

[3] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, Jun 2012, pp. 1–6.

[4] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," in *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Lecce, Italy, Aug 2017, pp. 1–6.

[5] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violent interaction detection in video based on deep learning," *Journal of Physics: Conf. Series*, vol. 844, p. 012044, Jun 2017.

[6] A. Hanson, K. PNVR, S. Krishnagopal, and L. Davis, "Bidirectional convolutional LSTM for the detection of violence in videos," in *European Conf. on Computer Vision - Workshops*, Munich, Germany, Sep 2018, pp. 280–295.

[7] A. Mumtaz, A. B. Sargano, and Z. Habib, "Violence detection in surveillance videos with deep network using transfer learning," in *European Conf. on Electrical Engineering and Computer Science*, Bern, Switzerland, Dec 2018, pp. 558–563.

[8] S. Accattoli, P. Sernani, N. Falcionelli, D. N. Mekuria, and A. F. Dragoni, "Violence detection in videos by combining 3d convolutional neural networks and support vector machines," *Appl. Artif. Intell.*, vol. 34, no. 4, pp. 329–344, 2020.

[9] T. Virtanen, M. D. Plumbley, and D. Ellis, Eds., *Computational Analysis of Sound Scenes and Events*, 1st ed. Springer Int. Publishing, 2018.

[10] W. Huang, T.-K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream detection for home applications," in *IEEE Conf. on Industrial Electronics and Applications*, Taichung, Taiwan, Jun 2010, pp. 2115–2120.

[11] J. Pohjalainen, T. Raitio, and P. Alku, "Detection of shouted speech in the presence of ambient noise," in *Interspeech*, Florence, Italy, Aug 2011, pp. 2621–2624.

[12] Y. Lee, D. Han, and H. Ko, "Acoustic signal based abnormal event detection in indoor environment using multiclass adaboost," *IEEE Trans. on Consumer Electronics*, vol. 59, pp. 615–622, Aug 2013.

[13] P. Ribeiro, R. Audigier, and Q. Pham, "Rimoc, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance," *Computer Vision and Image Understanding*, vol. 144, pp. 121–143, 2016.

[14] J.-L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *Intelligent Transportation Systems Conf.*, Sep 2006, pp. 733–738.

[15] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Shanghai, China, Mar 2016, pp. 6460–6464.

[16] P. Laffitte, Y. Wang, D. Sodoyer, and L. Girin, "Assessing the performances of different neural network architectures for the detection of screams and shouts in public transportation," *Expert Systems with Applications*, vol. 117, pp. 29–41, Mar 2019.

[17] T. Marteau, S. Afanou, D. Sodoyer, S. Ambellouis, and F. Boukour, "Audio events detection in noisy embedded railway environments," in *EDCC 2020, Workshop on Artificial Intelligence for RAILwayS)*, Munich, Germany, Sep 2020, pp. 20–30.

[18] Q. C. Pham, A. Lapeyronnie, C. Baudry, L. Lucat, P. Sayd, S. Ambellouis, D. Sodoyer, A. Flancquart, A. Barcelo, F. Heer, F. Ganansia, and V. Delcourt, "Audio-video surveillance system for public transportation," in *Int. Conf. on Image Process. Theory Tools and Applications*, Paris, France, Jul 2010, pp. 47–53.

[19] R. Zouaoui, R. Audigier, S. Ambellouis, F. Capman, H. Benhadda, S. Joudrier, D. Sodoyer, and T. Lamarque, "Embedded security system for multi-modal surveillance in a railway carriage," in *SPIE security and defence*, Toulouse, France, Sept 2015, p. Paper N9652–11.

[20] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Trans. on Pattern Analysis and Machine Intell.*, vol. 41, pp. 423–443, Feb 2019.

[21] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 28, pp. 1788–1800, 2020.

[22] B. T. Jin, L. Abdelrahman, C. K. Chen, and A. Khanzada, "Fusical: Multimodal fusion for video sentiment," in *Int. Conf. on Multimodal Interaction*, Virtual Event, The Netherlands, Oct 2020, pp. 798–806.

[23] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, "Variational bayesian inference for audio-visual tracking of multiple speakers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 1761–776, 2021.

[24] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to Look: Action Recognition by Previewing Audio," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Jun 2020.

[25] T. Heittola, A. Mesaros, and T. Virtanen, "Acoustic scene classification in DCASE 2020 challenge: Generalization across devices and low complexity solutions," in *Detection and Classification of Acoust. Scenes and Events - Workshops*, Tokyo, Japan (full virtual), Nov 2020, pp. 56–60.

[26] J. Cramer, H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, Brighton, United Kingdom, May 2019, pp. 3852–3856.

[27] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *IEEE Int. Conf. on Computer Vision*, Venice, Italy, Oct 2017, pp. 609–617.

[28] ——, "Objects that sound," in *European Conf. on Computer Vision*, Munich, Germany, Sep 2018, pp. 451–466.

[29] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "Audio-visual scene classification: Analysis of DCASE 2021 challenge submissions," in *Detection and Classification of Acoust. Scenes and Events 2021*, Online, Nov 2021, pp. 45–49.

[30] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, Jul 2017, pp. 4724–4733.

[31] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," May 2017.

[32] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *European Conf. on Computer Vision*, Munich, Germany, Sep 2018, pp. 252–268.

[33] J. Venn, "On the diagrammatic and mechanical representation of propositions and reasonings," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 10, no. 59, pp. 1–18, 1880.