CNNs and Fisher Vectors for No-Audio Multimodal Speech Detection

Jose Vargas[1], Hayley Hung[1]
[1]Delft University of Technology, Netherlands
j.d.vargasquiros@tudelft.nl,h.hung@tudelft.nl

ABSTRACT

This paper presents the algorithms that the organisers deployed for the automatic Behavior Analysis (HBA) task in MediaEval 2019, consisting on the detection of speech in social interaction from body-worn acceleration and video only. For acceleration-based prediction, a CNN with access to a window of 3s around and including the one-second prediction window is shown to perform remarkably. For video-based prediction, a Fisher vector pipeline with access only to the prediction window of 1s was found to perform significantly worse, while the late fusion of both approaches resulted in a small improvement.

1 INTRODUCTION

The No-Audio Multimodal Speech Detection task [5] of MediaEval 2019 aims to study the problem of determining the speaking status of standing subjects in crowded mingling scenarios. The non-verbal input consists in accelerometer readings from a wearable devices worn around the neck of the subjects, and video recorded from overhead cameras.

The problem is of interest because the automatic detection of speech from the visual modality allows for more detailed computational analyses of social behavior when audio of conversations is not available. The importance of the acceleration modality is twofold. First, the use of accelerometers in wearable devices poses little privacy concerns and such devices have therefore become common in social interaction datasets, providing limited but exploitable information about the body movement of subjects. Second, being a proxy for body movement, insights about how to best detect social actions from acceleration information could potentially transfer to other modalities like video.

2 RELATED WORK

Using the same dataset, a previous submission for the same task [1] makes use of PSD feature extraction and a Transductive Parameter Transfer method for classifying based on acceleration and dense trajectories and a Multiple Instance Learning method for classifying the video modality. Late fusion is also used and results in an increase in performance. Both methods were proposed in separate papers for the speech detection task [2, 6].

Research in psychology and computer science has investigated the synchrony between speech and gesture [4] and the role that gestures play in complementing or being redundant to speech [8, 9].

Very little literature is concerned with the specific task of recognizing speaking status without access to audio. Much more concern has received the automatic detection of gestures, possibly the most salient manifestation of speech behavior in our dataset. Although gesture recognition can certainly be treated as an action recognition or localization problem, it has received some attention in studies that focus specifically on this task [3, 10, 15, 16]. The datasets used, however, differ in that they normally offer a clear frontal view of a single person.

3 APPROACH

The task was approached using a traditional dense trajectories pipeline for video-based detection. For acceleration-based detection, a one-dimensional convolutional neural network with access to context outside of the prediction window was used. Multimodal detection was approached via late fusion of classification scores.

3.1 Estimation from video: Dense Trajectories and Fisher Vectors

The method for video classification was based on dense trajectories [13, 14] due to their relative simplicity and competitive performance even when compared with more recent deep learning approaches for action recognition.

Fisher vectors [12], and specially their improved variant , [11] were found to perform remarkably well in comparisson with Multiple Instance Learning [2] in experiments with 3-second windows, and were therefore chosen as classification algorithm.

Fisher vectors provide a way to obtain a compact feature vector from an arbitrary number of local features by making use of the additive property of log-likelihood in a generative model (see figure 2). Let $X = \{x_t, t = 1...T\}$ be the set of T local descriptors of dimensionality D extracted from an image and u_λ be the probability density function with parameters λ. The fisher score is defined as the gradient of the log-likehood over X, with respect to the model parameters:

$$G_\lambda^X = \frac{1}{T}\nabla_\lambda \log u_\lambda(X) \tag{1}$$

where λ denotes the model parameters. The fisher vector is a normalized version of the Fisher score:

$$\mathcal{G}_\lambda^X = L_\lambda G_\lambda^X \tag{2}$$

where normalization by L_λ corresponds to whitening of the dimensions. Any generative model can be used as u_λ. We chose a Gaussian mixture model (GMM) with K components with diagonal covariance matrices, in line with previous work [11]. The parameters λ of a GMM are $\lambda = \{w_i, \mu_i, \sigma_i^2, i = 1, \ldots, K\}$, where w_i, μ_i and σ_i^2 are the mixture weight, mean vector and diagonal of the covariance matrix of Gaussian i. Mean and standard deviation are the only parameters considered because mixture weights add little

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

additional information [11]. Under the assumption of independence of local descriptors:

$$G_\lambda^X = \frac{1}{T}\sum_{t=1}^{T} \nabla_\lambda \log u_\lambda(x_t) \tag{3}$$

Let $\gamma_t(i)$ be the soft assignment of descriptor x_t to Gaussian i:

$$\gamma_t(i) = \frac{w_i u_i(x_t)}{\sum_{j=1}^{K} w_j u_j(x_t)} \tag{4}$$

Derivation of the gradients leads to:

$$\mathcal{G}_{\mu,i}^X = \frac{1}{T\sqrt{w_i}}\sum_{t=1}^{T}\gamma_t(i)\left(\frac{x_t-\mu_i}{\sigma_i}\right) \tag{5}$$

$$\mathcal{G}_{\sigma,i}^X = \frac{1}{T\sqrt{2w_i}}\sum_{t=1}^{T}\gamma_t(i)\left[\frac{(x_t-\mu_i)^2}{\sigma_i^2}-1\right] \tag{6}$$

where the division between vectors is term-by-term. The Fisher Vector aggregates all gradients into a vector of $2KD$ dimensions. Finally Fisher vectors are normalized by dividing by their L2 norm and then power-normalized with $f(z) = sign(z)\sqrt{|z|}$.

For the task, person videos were resized to 100x100px. A set of 200 one-second windows were sampled per person, reducing the size of the training set to 10800 examples, due to the large size of the represenation. A GMM with 256 components was used. Fisher vectors were fed into a linear SVM classifier. 4-fold cross validation at the subject level was used to determine the optimal regularization parameter.

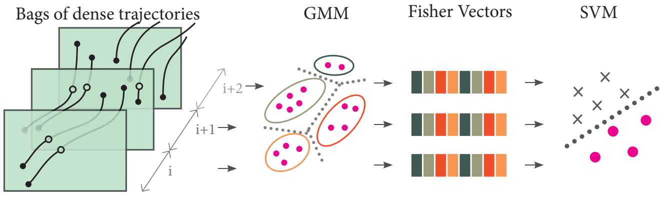

Figure 1: Fisher vectors pipeline.

3.2 Estimation from acceleration: 1-D Convolutional Neural Network (CNN)

For the classification of one-second windows using acceleration, a one-dimensional CNN was chosen. The architecture was based on the two-dimensional AlexNet [7]. The ratios between number of channels was preserved but the number of channels was reduced due to the reduced complexity of the input (see figure 2). Because experiments have revealed that 3-second windows are more informative for the detection of speaking status, the network was fed 3-second windows to give it access to a wider context, but only the middle second is predicted. The data was padded with zeros at both ends.

A sliding window of 3s with stride of 1s was used to produce the training examples. Data was pre-processed by z-score standardization on each axis, to reduce the effect of gravity and device miscalibration.

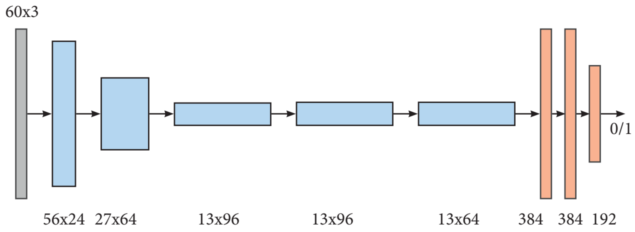

Figure 2: Architecture of the 1D-CNN used. Input data has 3 channels corresponding to axes X, Y and X of the accelerometer. Filter sizes are 5 for the first convolutional layer and 3 for the rest of the layers, with unit padding. As with AlexNet, first, second and last layers are followed by a max-pooling layer kernel size 3 and stride of 2.

3.3 Multimodal estimation: late fusion

Late fusion of the scores of both modalities was used to obtain multimodal scores, by training a logistic regressor with no regularization on the output scores of both modalities.

4 RESULTS AND ANALYSIS

Table 1 presents the results on the provided test set.

Submission	Method	AUC
This submission	1D CNN	0.692
	Fisher vectors	0.552
	Fusion	0.693
Past submission [1]	TPT	0.656
	MILES	0.549
	Fusion	0.658

Table 1: Test results.

5 DISCUSSION AND OUTLOOK

Although the submitted results indicate much better performance from the acceleration-based method, our experiments using prediction windows of 3s for both methods have resulted in very similar performance, indicating that the larger context fed into the CNN is useful for prediction. The experiments made for the submission suggested multiple areas for possible future work. One of them relates to how to give dense trajectory methods context in an equivalent way. Giving dense-trajectory-based methods access to context for high-resolution prediction is not straightforward given that aggregation methods like Fisher vectors are time-agnostic, unlike a CNN which only compresses its time dimension.

The comparison with the results of our past submission indicates that Fisher Vectors are capable of outperforming MILES. Our experiments also showed that personalisation using TPT does not deliver better results for this dataset, even when compared with a more simple Logistic Regressor.

ACKNOWLEDGMENTS

A special acknowledgement goes to Ekin Gedik and Laura Cabrera-Quiros for their support and input during the making of the paper. This task is supported by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606.

REFERENCES

[1] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. 2018. The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates. *IEEE Transactions on Affective Computing* (2018), 1–17. https://doi.org/10.1109/TAFFC.2018.2848914

[2] Laura Cabrera-Quiros, David M.J. Tax, and Hayley Hung. 2018. Gestures in-the-wild : detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration. (2018), 1–10.

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2016. Using Convolutional 3D Neural Networks for User-independent continuous gesture recognition. *Proceedings - International Conference on Pattern Recognition* 0 (2016), 49–54. https://doi.org/10.1109/ICPR.2016.7899606

[4] Anna Esposito and Antonietta M. Esposito. 2011. On speech and gestures synchrony. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6800 LNCS (2011), 252–272. https://doi.org/10.1007/978-3-642-25775-9_25

[5] Ekin Gedik, Laura Cabrera-Quiros, and Hayley Hung. 2019. No-Audio Multimodal Speech Detection task at MediaEval 2019. (2019).

[6] Ekin Gedik and Hayley Hung. 2017. Personalised models for speech detection from body movements using transductive parameter transfer. *Personal and Ubiquitous Computing* 21, 4 (2017), 723–737. https://doi.org/10.1007/s00779-017-1006-4

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS Proceedings* (2012). https://doi.org/10.1201/9781420010749

[8] Alex Lascarides and Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics* (2009). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.3741

[9] David McNeill. 1994. Hand and Mind: What Gestures Reveal About Thought. *University of Chicago Press* (1994). https://doi.org/10.1177/002383099403700208

[10] Qiguang Miao, Yunan Li, Wanli Ouyang, Zhenxin Ma, Xin Xu, Weikang Shi, and Xiaochun Cao. 2018. Multimodal Gesture Recognition Based on the ResC3D Network. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017* 2018-Janua (2018), 3047–3055. https://doi.org/10.1109/ICCVW.2017.360

[11] Florent Perronnin, Jorge Sanchez, and Thomas Mensink. 2010. Improving the Fisher Kernel for Large-Scale Image Classificatio. *ECCV 2010* (2010).

[12] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. 2013. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision* 105, 3 (2013), 222–245. https://doi.org/10.1007/s11263-013-0636-x

[13] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng Lin Liu. 2011. Action recognition by dense trajectories. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2011), 3169–3176. https://doi.org/10.1109/CVPR.2011.5995407 arXiv:1505.04868

[14] Heng Wang, Cordelia Schmid, Heng Wang, Cordelia Schmid, Action Recognition, Trajectories Iccv, Heng Wang, and Cordelia Schmid. 2013. Action Recognition with Improved Trajectories. *ICCV - IEEE International Conference on Computer Vision* December (2013), 3551–3558.

[15] Huogen Wang, Pichao Wang, Zhanjie Song, and Wanqing Li. 2018. Large-scale multimodal gesture recognition using heterogeneous networks. *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017* 2018-Janua (2018), 3129–3137. https://doi.org/10.1109/ICCVW.2017.370

[16] X Zabulis, H Baltzakis, and a Argyros. 2009. Vision-based hand gesture recognition for human-computer interaction. *The Universal Access ...* (2009), 1–56. http://users.ics.forth.gr/