Transductive Parameter Transfer, Bags of Dense Trajectories and MILES for No-Audio Multimodal Speech Detection

Laura Cabrera-Quiros[1,2], Ekin Gedik[1], Hayley Hung[1]
[1]Delft University of Technology, Netherlands
[2]Instituto Tecnológico de Costa Rica, Costa Rica.
{l.c.cabreraquiros,e.gedik,h.hung}@tudelft.nl

ABSTRACT

This paper presents the algorithms that task organisers deployed for the automatic Human Behaviour Analysis (HBA) task of the MediaEval 2018. HBA task aims to investigate alternate modalities of video and body-worn acceleration for the detection of speaking status. For unimodal estimation from acceleration, a transfer learning approach, Transductive Parameter Transfer (TPT), which is shown to perform satisfactorily in a similar setting[4] is employed. For the estimation from the video modality, bags of Dense Trajectories were used in a multiple instance learning approach (MILES) [2]. Finally, late fusion is used for combining the outputs from both modalities. The multi-modal approach resulted in a mean AUC of 0.658, outperforming the performance of both single modality approaches.

1 INTRODUCTION

The Human Behaviour Analysis (HBA) task of MediaEval 2018 focuses on non-audio speaking status detection in crowded mingling events [1]. Such events are interesting since they are concentrated moments for people to interact freely, resulting in unstructured and varied social behaviour. Since speaking turns are shown to be vital units of social behaviour [9], their automatic detection makes detailed analysis of social behaviour possible.

Traditionally, audio is used for the detection of speech. However, the dense nature of large gatherings introduces restrictions such as background noise, making the use of audio challenging. In order to overcome this challenge, the HBA task investigates the alternative modalities of wearable acceleration and video for the detection of speaking status. The main idea behind this approach is backed by prior work in social science where speakers were shown to move (e.g. gesture) during speech [5].

The task requires participants to provide solutions for unimodal estimations, both for acceleration and video, and a multimodal estimation. For more details about the task, please refer to [1].

For acceleration, we employed the transfer learning method called Transductive Parameter Transfer (TPT) which was shown to perform satisfactorily in a similar setting [4]. Speaker estimation from video is carried out by extracting bags of dense trajectories and using MILES (a multiple instance learning method) for classification. This approach from video allow us to overcome the cross-contamination of subjects standing close together due to their respective overlapping bounding boxes. Finally, the multimodal estimation is done by combining the outputs of these two unimodal classification approaches using late fusion. In the following section, we will explain these approaches in detail.

2 METHODOLOGY

2.1 Estimation from acceleration: TPT

Even though speakers are known to act differently from non-speakers [5], their behaviours vary greatly, making automatic estimation from acceleration a challenging task. In order to account for this variance, we employed a transfer learning model called TPT which can provide personalised models. It computes the parameters of the optimal classifier for a target dataset X^t given a set of source datasets with their own corresponding optimal classifiers. The classifier for the target data is computed without using any label information for the target dataset. The method was first proposed for facial expression detection [7]. A specialised version tuned for speaking status detection from acceleration was presented in [4].

Let N source datasets with label information and the unlabelled target dataset be defined as $D^s_1, ..., D^s_N$, $D^s_i = \left\{x^s_j, y^s_j\right\}_{j=1}^{n^s_i}$ and $X^t = \{x^t_j\}_{j=1}^{n_t}$, the following steps are taken for computing the optimal parameters $(\boldsymbol{w_t}, c_t)$ for X^t (where w and c correspond to regression coefficients and the intercept, respectively):

(1) $\{\boldsymbol{\theta_i} = (\boldsymbol{w_i}, c_i)\}_{i=1}^{N}$ is computed using L2 penalized logistic regression,
(2) Training set $\tau = \{X^s_i, \boldsymbol{\theta_i}\}_{i=1}^{N}$ is created,
(3) The kernel matrix $\boldsymbol{K}$ that defines the distances between distributions where $\boldsymbol{K_{ij}} = \kappa(X^s_i, X^s_j)$ is computed with an Earth Mover's distance kernel [6].
(4) Given $\boldsymbol{K}$ and τ, $\hat{f}(.)$, the mapping between marginal distributions of the datasets and their optimal parameters, is computed with Kernel Ridge Regression.
(5) $(\boldsymbol{w_t}, c_t) = \hat{f}(X^t)$is computed using the mapping obtained in the former step.

For a more detailed explanation of each step, readers can refer to [4]. We used statistical and spectral features extracted from 3s windows with 1.5s overlap for each axis of the raw acceleration signal, absolute values of the acceleration signal and the magnitude of the acceleration. As the statistical features, mean and variance values are calculated. The power spectral density computed using 8 bins with logarithmic spacing forms our spectral feature set. Each axis of the acceleration is standardised to have zero mean and unit variance. The probability outputs are then upsampled to 1s windows.

2.2 Estimation from video: Bags of dense trajectories and MILES

The video for this problem is inherently noisy, as we can have more than one person in the video for our person of interest (eg. people talking close together). Thus, we propose to use bags of dense trajectories to overcome the cross-contamination in the video.

MediaEval'18, 29-31 October 2018, France

First, we extract the dense trajectories for all the participants using the method proposed by Wang et.al. [10]. Then, these trajectories are clustered into bags using a sliding window of 3sec with an overlap of 1.5sec. Thus, all the trajectories that overlap at least an 80% with the window are part of the bag for this window.

This clustering into bags results in a set $\mathbf{B}^s$ of bags (positive and negative) for subject s, where $s = \{1, ..., S\}$ and S is the total number of subjects. A bag from this set is then $\mathbf{B}^s_j$, where $j = \{1..., N^s\}$, and N^s is the total number of bags possible for subject s. Moreover, we cluster also in space the trajectories within a bag using k-means clustering. We do so to account for spatial similarities and for computational efficiency. This way, the trajectories for each bag are clustered into the k most representative prototypes for the bag.

Note that each bag $\mathbf{B}^s_j$ will consist of *good* trajectories (corresponding to the subject s) and *bad or noise* trajectories (other subjects or shadows and other background artifacts). Thus, we need to treat the samples in a bag differently, instead of each trajectory independently. This is the main motivation for using a Multiple Instance Learning (MIL) approach for classification on video.

As our MIL approach we use Multiple Instance Learning via Embedded Instance Selection (MILES)[2]. Overall, MILES classifies a bag by considering both contributing information (e.g. trajectories of subject s in our case) and opposing information (e.g. trajectories from other subjects or background). It does so by creating a *concept* in an embedded space and comparing all instances to this concept.

Let us define $\mathbf{B} = \{\mathbf{B}^1, \mathbf{B}^2, ..., \mathbf{B}^S\}$, as the set of bags for all participants in the training set. $\mathbf{B}_a$ is then a bag of this set $\mathbf{B}$, where $a = \{1..., A\}$ and A is the sum of the total number of bags for all S subjects. $\mathbf{x}^j_a$ is then an instance (prototype trajectory) from this bag. For a given bag $\mathbf{B}_a$ the measure of similarity between this bag and all other instances (disregarding their bag) is calculated by

$$s(\mathbf{x}^k, \mathbf{B}_a) = \max_b \exp\left(-\frac{||x_{ab} - \mathbf{x}^k||^2}{\sigma^2}\right) \qquad (1)$$

where $\mathbf{x}^k$ is the set of instances in the training and x_{ab} is a given instance b within bag $\mathbf{B}_a$. Thus, bag $\mathbf{B}_a$ is embedded into a space of similarities defined as

$$\mathbf{m}(\mathbf{B}_a) = [s(\mathbf{x}^1, \mathbf{B}_a), s(\mathbf{x}^2, \mathbf{B}_a), ..., s(\mathbf{x}^{n_a}, \mathbf{B}_a)]^T \qquad (2)$$

where n_a is the total number of instances in the training set. This results in the matrix representation of all training bags in the embedded space $(\mathrm{IF}_c) : \mathbf{m}(\mathbf{B}) = [\mathbf{m}(\mathbf{B}_1), ..., \mathbf{m}(\mathbf{B}_A)]$.

On this representation a (sparse) linear classifier is then trained. The classification of new bags is done by:

$$y = \mathrm{sign}(\sum_{k \in I} w^*_k s(\mathbf{x}^k, \mathbf{B}_{new}) + b^*) \qquad (3)$$

where I is the subset of instances with non-zero weights ($I = \{k : |w^*_k| > 0\}$). Note that instances with contributing information will have positive weights w^*_k, while those with opposing information will have negative weights. We used the MILES implementation in PRTools [3]. For more details, please refer to [2].

2.3 Multimodal estimation: Late fusion

After computing 1 second estimations from acceleration and video modalities with aforementioned methods, we combine the predictions of both methods using mean fusion [8]. If the video of the current subject is missing, we directly use the output of the TPT.

	Accel	Video	Fusion
Mean AUC±Std	0.656 ± 0.074	0.549 ± 0.079	0.658 ± 0.073

Table 1: Performances of each modality and their (late) fusion.

3 RESULTS

Table 1 presents the performances for each task. Similarly, we present the performance obtained for each participant in Figure 1. For unimodal estimations, mean AUC scores of 0.656 and 0.549 with standard deviations of 0.074 and 0.079 are obtained for acceleration and video. As it can be seen from the Figure 1, performance per participant is highly varied. This further supports the claim that the movement patterns of speakers are highly varied, making detection harder for some than others.

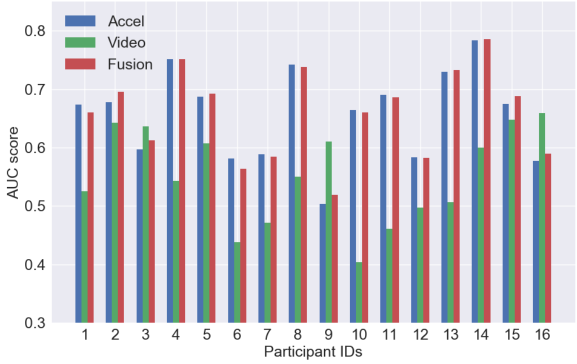

Figure 1: Performances per participant (p. independent)

Relatively low performance of the video modality is probably caused by the missing video data for some participants. These missing intervals are included in the performance evaluation dropping the overall performance for that participant. Cases where acceleration modality are outperformed by video further show the multimodal nature of the problem.

Moreover, the data present from the video can be noisy due to occlusions between the participants. Our MIL approach for video could tackle this problem up to a certain degree, but some cases are too crowded to be tackled from the video alone.

Finally, we can see that even with a basic fusion technique like mean fusion, a multimodal approach provided better performance than the single modalities. Even though the overall performance difference is marginal, mean fusion guaranteed similar or higher performance scores than both modalities. We argue that with a more sophisticated fusion approach, it should be possible to exploit the multimodal nature of the problem even more. A possible direction of research is addressing the occlusion segments during video in a smart fusion manner.

4 CONCLUSION

In this paper, we presented our approach for no-audio speech detection. The promising performances showed the possibility of tackling such a challenging task. Highest performance scores obtained by the multimodal fusion further supported the multimodal nature of the problem. However, there is still a huge room for improvement. We believe with the help of many, it will be possible to finally solve this challenging problem.

ACKNOWLEDGMENTS

This task is partially supported by the Instituto Tecnológico de Costa Rica and the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606.

REFERENCES

[1] L. Cabrera-Quiros, E. Gedik, and H. Hung. 2018. No-Audio Multimodal Speech Detection in Crowded Social Settings task at MediaEval 2018. *MediaEval* (2018).

[2] Y. Chen, J. Bi, and J.Z. Wang. 2006. MILES: Multiple-Instance Learning via Embedded Instance Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2006).

[3] P. Duin, R.P.W. Juszcak, P. Paclik, E. Pekalska, D. de Ridder, and D.M.J. Tax. 2017. PRTools, A Matlab Toolbox for Pattern Recognition. (March 2017). version 5.3.

[4] 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.

[5] David McNeill. 2000. *Language and gesture*. Vol. 2. Cambridge University Press.

[6] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40, 2 (2000), 99–121.

[7] Enver Sangineto, Gloria Zen, Elisa Ricci, and Nicu Sebe. 2014. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. In *Proceedings of the ACM international conference on multimedia*. ACM, 357–366.

[8] David MJ Tax, Martijn Van Breukelen, Robert PW Duin, and Josef Kittler. 2000. Combining multiple classifiers by averaging or by multiplying? *Pattern recognition* 33, 9 (2000), 1475–1485.

[9] Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3, 1 (2012), 69–87.

[10] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. 2013. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *Intern. Journal of Computer Vision* (2013).