Extracting audio-visual features for emotion recognition through active feature selection

EXTRACTING AUDIO-VISUAL FEATURES FOR EMOTION RECOGNITION THROUGH ACTIVE FEATURE SELECTION

Fasih Haider, Senja Pollak, Pierre Albert and Saturnino Luz

Usher Institute, Edinburgh Medical School, The University of Edinburgh, Edinburgh, UK

ABSTRACT

Research in automatic emotion recognition has seldom addressed the issue of computational resource utilisation. With the advent of ambient technology, which employs a variety of low-power, resource constrained devices, this issue is increasingly gaining interest. This is especially the case in the context of health and elderly care technologies, where interventions aim at maintaining the user's independence as unobtrusively as possible. In this context, efforts are being made to model human social signals such as affects using low-cost technologies, which can aid health monitoring. This paper presents an Active Feature Selection (AFS) method using self-organized maps neural networks for emotion recognition in the wild. The AFS is used for feature subsets selection from three different feature sets: 62 out of 88 features were selected for *eGeMAPs*, 21 out of 988 for *emobase*, and 140 out of 2832 for *LBPTOP* features. The results show that the features subsets selected by AFS provide better results than the entire feature set and PCA dimensionality reduction method. The best improvement is observed on *emobase* features, followed by *eGeMAPs*. For visual features, nearly the same results are obtained with a significant reduction in dimensionality (only 5% of the full feature set is required for the same level of accuracy). The weighted score fusion results in an improvement, leading to 43.40% and 40.12% accuracies on the EmotiW 2018 validation and test datasets respectively.

Index Terms— Feature Engineering, Feature Transformation, Feature Extraction, Feature selection, Emotion Recognition, Affective Computing

I. INTRODUCTION

The emerging fields of social signal processing and affective computing seek to build models to automatically characterise human behaviours in interactive situations. This includes the detection of emotions and attitudes which can, among other things, influence communication effectiveness both in dialogue and in presentations. Methods developed in these fields have found applications in the analysis of clinician-patient communication [26], education [22], entertainment [4] and cognitive health monitoring [6]. In the SAAM project [6], we are employing Ambient Assisted Living (AAL) technologies to analyse activities and health status, and provide personalised multimodal coaching to elderly persons living on their own or in assisted care settings. Such activities and status include mobility, sleep, social activity, air quality, cardiovascular health, diet [15] and attitudes [10].

Audio-visual signals are used in a number of automatic prediction tasks, including cognitive state detection [3], presentation skills assessment [11, 13] and emotion recognition [10, 12, 8, 9, 14], the latter being also the topic of the audio-video challenge of the Emotion Recognition in the Wild Challenge (EmotiW 2018) [5] that we address in this paper. The approaches to the audio-visual signal analysis have employed very high-dimensional feature-space consisting of large numbers of potentially relevant acoustic/visual features. For audio signal, these features are usually obtained by applying statistical functionals to basic, energy, spectral and voicing related acoustic descriptors [9] extracted from speech intervals lasting a few seconds [27]. Although there is no general consensus on what the ideal set of descriptors should be, this "brute-force" approach of employing a large set of acoustic descriptors seems to outperform alternative (Markovian) approaches of modeling temporal dynamics on the classifier level [32]. On the other hand, the use of such high-dimensional datasets poses serious challenges for prediction (one of the facets of the so called "curse of dimensionality"). Higher-order statistical-functionals with typically a high degree of redundancy on the feature set, as well as features of poor descriptive value. Dimensionality reduction is therefore a vital part of automatic classification in these types of datasets. Moreover, such high-dimensional approaches are not suitable for designing an embodied emotion recognition system with low power, cost, memory and computational resources such as using Raspberry Pi Zero[1].

This study extends our previous work [15, 16] and the main contribution of this paper is to demonstrate the performance of 'Active Feature Selection (AFS)' method (which

This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 769661, SAAM project.

[1]https://www.raspberrypi.org/products/raspberry-pi-zero/ (last accessed June 2019)

978-1-7281-2723-1/19/$31.00 ©2019 IEEE

is recently proposed and tested only on speech features for eating conditions recognition [15] and emotion recognition [16]) along with Principle Component Analyses (PCA) on speech and visual features. We also propose an ensemble method for fusion and demonstrate its results for emotion recognition in the wild. This study is the first demonstration of AFS method on visual features and on a larger corpora in-terms of number of subjects than previous studies [15, 16].

II. BACKGROUND AND RELATED WORK

The emotion recognition in the wild data set [5] have been extensively used in the literature and the best performing approaches achieve the accuracy around 60% [18, 24]. For example, Hu et al. [18] proposes a deep CNN architecture, where a Supervised Scoring Ensemble (SSE) method is used for dense supervision to diverse feature layers (not only deep layers, but also to intermediate layers and shallow layers). The visual models are mainly based on ResNet, DenseNet and HoloNet, where their SSE learning contributes to achieving much higher accuracy compared to standard training; in addition they add a baseline hand-crafted model by Yao et al. [34]. For audio model, they use openSMILE [9] features for SVM classifier. When setting the focus on audio models only, some studies do not report separate mono-modal results. Numerous participants [28, 18] extract audio features with OpenSMILE using *de facto* standard presets: IS10, GeMAPS, eGeMAPS. Vielzeuf et al. achieve 36.5% accuracy [28]. They use a two-layer perceptron to predict classes as well as compact descriptors from these features [28]. Wang et al. [31] used both IS10 and MFCC features, achieving respectively 38% and 39.5% accuracy.

Based on the above literature we have concluded that the feature dimensionality is very high (in some cases it is near 10k) for the classification task. Although the accuracy is promising (around 60%) no effort is spent on dimensionality reduction (removing noisy/redundant features) to reduce the 'curse of dimensionality' and computational resources (i.e. extraction of a subset of feature set instead of whole feature set results in reduction of usage of machine memory, cost, computational resources and power). There are many dimensionality reduction methods: supervised methods, such as correlation based feature selection [17], require labelled data, while unsupervised approaches, such as PCA [1] and independent component analysis (ICA) [29], do not require labeled datasets. Recently, the efforts have been spent to reduce the dimensionality using PCA to improve the results for emotion recognition from speech [20, 2, 30] in different settings such as noisy setting [2] but not in the wild setting. In this study, we demonstrated the PCA performance over three different feature sets (i.e. emobase, eGeMAPs and LBPTOP) and compare the results to our recently developed AFS [15] and fusion method.

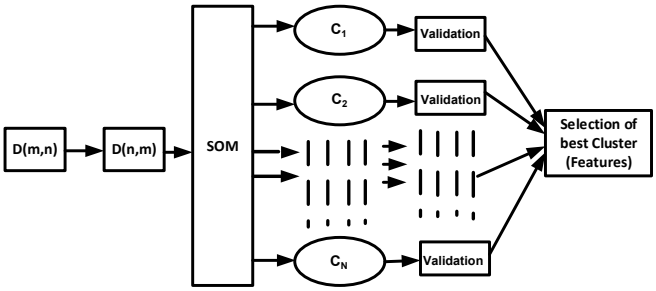

Fig. 1. Active feature selection method: D(m,n) represents the data where *m* is the total number of training instances (773 training data instances in EMOTIW challenge) and *n* is the total number of dimensions (988 *emobase*, 88 for *eGeMAPs* and 2832 for LBPTOP).

III. THE ACTIVE FEATURE SELECTION METHOD

In this section, we describe our 'Active Feature Selection' (AFS) method which divides a feature set into subsets. It involves clustering the data-set into N (where $N = 5, 10, 15, ..., 100$) clusters using self-organizing maps (with 200 iterations and batch training) [21], and then evaluating discrimination power of features present in each cluster C_N using validation dataset, as depicted in Figure 1, and selecting the cluster with the highest validation accuracy. Here, we are not clustering the number of instances but the dimensions and not evaluating each feature separately but evaluating all the features in one cluster together. Our hypothesis is that the noisy features have different characteristics than informative features, and that clustering the features will divide the features into many subsets according to their common characteristics. An example of self-organizing clustering is depicted in Figure 2, where 2832 features ($LBPTOP$) are clustered into 10 clusters, where the features present in cluster number 9 (140 out of 2832) provide better results (accuracy on the validation dataset) than features in other clusters. The distance between these clusters is depicted in Figure 3.

IV. EXPERIMENTATION

IV-A. Data Set

The EmotiW6 AFEW data set [5] consists of video abstracts from movies and TV shows, labelled with the traditional set of 6+1 emotions formalized by Ekman [7]: Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. Both training and validation sets are imbalanced between emotions with respectively 133, 74, 81, 150, 144, 117, 74 and 64, 40, 46, 63, 63, 61, 46 sequences each. Detailed statistics for sequences and meta-data are summarized in Table I. The test set contains 653 videos.

Table I. Statistics of the data set

Training set				
#sequences	773			
#persons	229	female: 89	male: 140	
	min	max	mean	median
age	5	76	35	34
length	320	5882	2458	2287
Validation set				
#sequences	383			
#persons	135	female: 55	male: 80	
	min	max	mean	median
age	10	70	35	35
length	500	6121	2263	2082

IV-B. Feature Extraction

We use the openSMILE [9] toolkit for acoustic feature extraction: in total we extract 988 *emobase* and 88 *eGeMAPS* features (using $emobase.config$ and $eGeMAPSv01a.conf$ configuration files), which have been widely used for emotion recognition in the past [23, 9]. In addition, we used the visual features, namely the 2832 LBPTOP features [5].

In our experiments, we use three different feature set types, of different dimensions: the entire feature sets (All), the feature sets after the proposed AFS method (described in Section III) and the PCA feature sets, where PCA was chosen as an alternative dimensionality reduction method for comparison.

Feature sets - All: Full *emobase* feature set (988 acoustic features), *eGeMAPS* feature set (88 acoustic features) and LBPTOP feature set (2832 visual features).

Feature sets - PCA: transformed feature set of *emobase*, *eGeMAPs* and LBPTOP features using PCA. Then using sequential forward selection to select the number of dimensions for classification. We start with an empty feature set and keep increasing feature set by adding PCA dimensions one by one for classification task. We select the number of dimensions of PCA for classification which provides the best results on the validation set (i.e. validation accuracy is the selection criteria).

Feature sets - AFS: Feature sub-sets selection from the three feature sets (*emobase*, *eGeMAPs* and LBPTOP) using AFS as detailed in Section III.

IV-C. Ensemble for Fusion

In addition, we propose the score fusion of feature sets using an ensemble method as described below. Score Fusion of classification output using the weighted fusion (**F**) method (for an ensemble of n classifiers) as shown in Equation *1*, where w_k is the assigned weight, with $w_1{+}w_2{+}...{+}w_n = 1$, and $y(k)$ is the k^{th} classifier output score. For attributing the weights, we used the Equations *2, 3* and *4* with $t = |\delta| : 0.001 : 1$ and $-0.05 < \delta < +0.05$ and selected those values for weights (w_k) that provide the best validation result. For score fusion of ensemble with two classifiers, the $y(3) = 0$, $\delta = 0$, $t = t/2$ and the Equation *4* is ignored. The weights for score fusion are set in Equations *5–13*. For example, the score fusion of two classifier outputs ($F_{AFS_{Emobase}+AFS_{LBPTOP}}$) has a weights of 0.714 (w_1) for $AFS_{Emobase}$ (LDA score of active feature selection using *emobase* feature) and 0.286 (w_2) for AFS_{LBPTOP} (LDA score of active feature selection using LBPTOP features).

IV-D. Classification Method

The classification is performed using Linear Discriminant Analysis (LDA). This classifier is employed in MATLAB[2] using the statistics and machine learning toolbox. LDA works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix [25]).

V. RESULT AND DISCUSSION

The classification is performed in training, validation and testing setting using the feature vectors described in Section IV-B and validation results are depicted in Table II. It is observed that the AFS method provides the best results for audio feature sets. However, it is unable to outperform the LBPTOP full feature set and results in almost of same accuracy but using very less dimension of features. These results also validated our hypothesis that by clustering the features we can remove noisy/redundant features, as by clustering the total number of features, we find a subset of features (62 out of 88 for eGeMAPs, 21 out of 988 for emobase and 140 out of 2832 for LBPTOP feature sets) which provide better/almost equal results than the full feature set and PCA feature set.

The AFS of *emobase* feature set (31.07% with a dimensionality of 21) provides more accurate results than AFS of *eGeMAPs* feature set (30.29% with a dimensionality of 62), suggesting that the combination of both features set may improve the classification performance. The LBPTOP feature set provides the best results for all three feature dimensionality (ALL (35.31%), PCA (29.11%) and AFS (35.04%)) suggesting that the visual information is more discriminative than audio information. The weighted score fusion using ensemble method results are depicted in Table III. Where *'AFS score fusion'* (39.89%) of *emobase*, *eGeMAPs* and LBPTOP features sets provides better results than *'All score Fusion'* (36.93%) and *'PCA score fusion'* (31.00%). By fusing the classifier score (ensemble method) of the features sets (full feature set, PCA and AFS feature sets) results in an improvement (43.40% on validation data and 40.12% on test data). The presented work is not able to outperform the previous methods [28, 18, 24] but the objective of this study is not to propose an emotion recognition system but to demonstrate the AFS and PCA methods performance over

[2]http://uk.mathworks.com/products/matlab/ (last accessed June 2019)

$$Fusion_{Score} = w_1.y(1) + w_2.y(2) + ... + w_n.y(n) \quad (1)$$

$$Fusion_1 = y(1).(1 - 2.t) + y(2).(t - \delta) + y(3).(t + \delta) \quad (2)$$

$$Fusion_2 = y(1).(t - \delta) + y(2).(1 - 2.t) + y(3).(t + \delta) \quad (3)$$

$$Fusion_3 = y(1).(t - \delta) + y(2).(t + \delta) + y(3).(1 - 2.t) \quad (4)$$

$$F_{AFS_{Emobase\ +\ AFS_{LBPTOP}}} = 0.714.AFS_{Emobase} + 0.286.AFS_{LBPTOP} \quad (5)$$

$$F_{AFSEmobase\ +\ AFSeGeMAPs} = 0.734.AFS_{Emobase} + 0.266.AFS_{eGeMAPs} \quad (6)$$

$$F_{ALL} = 0.116.ALL_{Emobase} + 0.392.ALL_{eGeMAPs} + 0.492.ALL_{LBPTOP} \quad (7)$$

$$F_{PCA} = 0.014.PCA_{Emobase} + 0.443.PCA_{eGeMAPs} + 0.543.PCA_{LBPTOP} \quad (8)$$

$$F_{AFS} = 0.592.AFS_{Emobase} + 0.154.AFS_{eGeMAPs} + 0.254.AFS_{LBPTOP} \quad (9)$$

$$F_{eGeMAPs} = 0.02.ALL_{eGeMAPs} + 0.02.PCA_{eGeMAPs} + 0.958.AFS_{eGeMAPs} \quad (10)$$

$$F_{Emobase} = 0.016.ALL_{Emobase} + 0.016.PCA_{Emobase} + 0.968.AFS_{Emobase} \quad (11)$$

$$F_{LBPTOP} = 0.3.ALL_{LBPTOP} + 0.4.PCA_{LBPTOP} + 0.3.AFS_{LBPTOP} \quad (12)$$

$$F_{Final} = 0.171.F_{eGeMAPs} + 0.271.F_{Emobase} + 0.558.F_{LBPTOP} \quad (13)$$

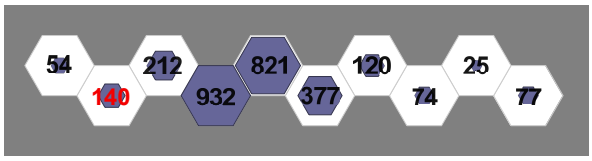

Fig. 2. Figure indicates the number of features present in each cluster (hexagon). Where N = 10 and the cluster with highest accuracy contain 140 LBPTOP features.

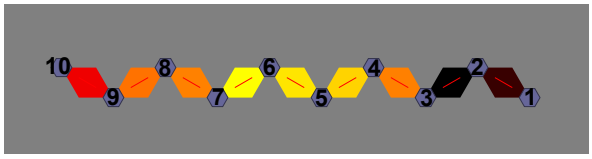

Fig. 3. Figure indicates the distance between clusters (darker color indicates greater distance between clusters). Where Cluster number 9 provides the best validation results for LBPTOP features

emotion recognition in the wild. In previous study [15], we demonstrate that the AFS method is able to select a feature sub-set (only audio features) which provide better results that entire feature set and PCA feature set for eating condition recognition. This study deploys AFS method on audio and visual features for emotion recognition task and a step towards in demonstrating the generalizability of AFS method. We demonstrate that the AFS method can select a feature sub-set which improves the emotion recognition performance over full feature set and PCA (with sequential forward selection of PCA dimensions) feature set. The AFS evaluates a full feature sub-set (clusters of feature set) instead of evaluating each feature separately and can help in the dimensionality reduction. One of the limitation of this study is the use of only LDA classifier and with other classifiers the results may change. That's why, the selected sub-sets of features using AFS should be tested with other advanced classifiers such as extreme learning machines [19] and partial least squares [33] along with other dimensionality reduction methods. Another limitation of this study is the evaluation of AFS on only two acoustic and one facial feature set. The performance of AFS may change when applied on other feature sets such as genetics features. The main strength of this study is the demonstration of AFS method on a relatively larger dataset in terms of subjects than previous studies [15, 16].

Table II. Best results for each experiment on validation (in %) dataset: Where *Accu.* is accuracy and *Dim.* is the number of dimensions used for obtaining the accuracy.

	Audio-eGeMAPS		Audio-emobase		Visual-LBPTOP	
	Accu.	Dim.	Accu.	Dim.	Accu.	Dim.
All	28.98	88	15.40	988	**35.31**	2832
PCA	27.15	77	28.20	33	29.11	60
AFS	**30.29**	62	**31.07**	21	35.04	140

Table III. Results on test data for AFS and weighted score fusion results on the validation dataset (in %).

Feature	Val (Accuracy)	Test (Accuracy)
Blind Guess	14.28	14.28
AFS-Emobase	31.07	26.49
AFS-LBPTOP	35.04	29.71
$F_{AFS_{Emobase}+AFS_{LBPTOP}}$	39.08	33.84
$F_{AFS_{Emobase}+AFS_{eGeMAPS}}$	33.94	–
F_{ALL}	36.93	–
F_{PCA}	31.00	–
F_{AFS}	**39.89**	36.14
$F_{eGeMAPS}$	30.81	–
$F_{emobase}$	32.11	–
F_{LBPTOP}	**39.89**	–
F_{Final}	***43.40***	40.12

VI. CONCLUSION

The results of Active Feature Selection (AFS) method has been demonstrated for emotion recognition in the wild. The subset of selected fetures using the AFS outperformed the full feature set and the PCA transformation (with sequential forward selection) for audio features using Linear Discrimination Analysis (LDA) classifier. However, for the visual features the AFS feature sub-set and full feature set provides almost the same results but with a significant reduction in the feature set to obtained the same accuracy. By fusing the classifier score of the features sets (full feature set, PCA and AFS feature sets) results in an improvement (43.40% on validation data and 40.12% on test data). In future we intend to evaluate the performance of the AFS method against other dimensionality reduction methods for multiple tasks and feature sets.

REFERENCES

[1] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[2] P. K. Aher, S. D. Daphal, and A. N. Cheeran, "Analysis of feature extraction techniques for improved emotion recognition in presence of additive noise," in *Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on*. IEEE, 2016, pp. 350–354.

[3] H. Akira, F. Haider, L. Cerrato, N. Campbell, and S. Luz, "Detection of cognitive states and their correlation to speech recognition performance in speech-to-speech machine translation systems," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 2539–2543.

[4] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: predicting personality from facial expressions of emotion in online conversational video," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 53–56.

[5] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *ACM International Conference on Multimodal Interaction*. ACM, 2018.

[6] Y. Dimitrov, Z. Gospodinova, M. Žnidaršič, B. Ženko, V. Veleva, and N. Miteva, "Social activity modelling and multimodal coaching for active aging," in *Procs. of Personalized Coaching for the Wellbeing of an Ageing Society, COACH'2019*, 2019.

[7] P. EKMAN, "Pictures of Facial Affect," *Consulting Psychologists Press*, 1976.

[8] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[9] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[10] F. Haider and S. Luz, "Attitude recognition using multi-resolution cochleagram features," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 3737–3741.

[11] F. Haider, L. Cerrato, N. Campbell, and S. Luz, "Presentation quality assessment using acoustic information and hand movements," in *Proceeding of 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[12] F. Haider, L. S. Cerrato, S. Luz, and N. Campbell, "Attitude recognition of video bloggers using audio-visual descriptors," in *Proceedings of the Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*, ser. MA3HMI '16. New York, NY, USA: ACM, 2016, pp. 38–42.

[13] F. Haider, F. A. Salim, S. Luz, C. Vogel, O. Conlan, and N. Campbell, "Visual, laughter, applause and spoken expression features for predicting engagement within ted talks," in *Proc. Interspeech 2017*, 2017, pp. 2381–2385.

[14] F. Haider, F. A. Salim, O. Conlan, and S. Luz, "An active feature transformation method for attitude recognition of video bloggers," in *Proc. Interspeech 2018*, 2018, pp. 431–435.

[15] F. Haider, S. Pollak, E. Zarogianni, and S. Luz, "SAAMEAT: active feature transformation and selection methods for the recognition of user eating conditions," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 564–568.

[16] F. Haider, S. Pollak, P. Albert, and S. Luz, "Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods," *arXiv:1908.10623*, 2019.

[17] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[18] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI 2017. New York, NY, USA: ACM, 2017, pp. 553–560.

[19] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2012.

[20] N. P. Jagini and R. R. Rao, "Exploring emotion specific features for emotion recognition system using pca approach," in *Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on*. IEEE, 2017, pp. 58–62.

[21] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1-3, pp. 1–6, 1998.

[22] C. Liu, R. A. Calvo, and R. Lim, "Improving medical students' awareness of their non-verbal communication through automated non-verbal behavior feedback," *Digital Education*, p. 11, 2016.

[23] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 494–501.

[24] C. Lu, W. Zheng, C. Li, C. Tang, S. Liu, S. Yan, and Y. Zong, "Multiple spatio-temporal feature learning for video-based emotion recognition in the wild," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 646–652.

[25] S. Raudys and R. P. W. Duin, "Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix," *Pattern Recognition Letters*, vol. 19, no. 5-6, pp. 385–392, Apr. 1998.

[26] P. Ryan, S. Luz, P. Albert, C. Vogel, C. Normand, and G. Elwyn, "Using artificial intelligence to assess clinicians' communication skills," *BMJ*, vol. 364, 2019.

[27] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[28] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI 2017. ACM, 2017, p. 569576.

[29] J. Wang and C.-I. Chang, "Independent component analysis-based dimensionality reduction with applications in hyperspectral image analysis," *IEEE transactions on geoscience and remote sensing*, vol. 44, no. 6, pp. 1586–1600, 2006.

[30] S. Wang, X. Ling, F. Zhang, and J. Tong, "Speech emotion recognition based on principal component analysis and back propagation neural network," in *Measuring Technology and Mechatronics Automation (ICMTMA), 2010 International Conference on*, vol. 3. IEEE, 2010, pp. 437–440.

[31] S. Wang, W. Wang, J. Zhao, S. Chen, Q. Jin, S. Zhang, and Y. Qin, "Emotion recognition with multimodal features and temporal models," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI 2017. New York, NY, USA: ACM, 2017, pp. 598–602.

[32] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music, and Sound have in Common," *Frontiers in Psychology*, vol. 4, May 2013.

[33] H. Wold, "Partial least squares," *Encyclopedia of statistical sciences*, 1985.

[34] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 451–458.