Towards a Better Gold Standard: Denoising and Modelling Continuous Emotion Annotations Based on Feature Agglomeration and Outlier Regularisation

WANG, Chen, *et al*.

Towards a Better Gold Standard: Denoising and Modelling Continuous Emotion Annotations Based on Feature Agglomeration and Outlier Regularisation

Chen Wang
University of Geneva
Geneva, Switzerland
chen.wang@unige.ch

Phil Lopes
University of Geneva
Geneva, Switzerland
phil.lopes@unige.ch

Thierry Pun
University of Geneva
Geneva, Switzerland
thierry.pun@unige.ch

Guillaume Chanel
University of Geneva
Geneva, Switzerland
guillaume.chanel@unige.ch

ABSTRACT

Emotions are often perceived by humans through a series of multimodal cues, such as verbal expressions, facial expressions and gestures. In order to recognise emotions automatically, reliable emotional labels are required to learn a mapping from human expressions to corresponding emotions. Dimensional emotion models have become popular and have been widely applied for annotating emotions continuously in the time domain. However, the statistical relationship between emotional dimensions is rarely studied. This paper provides a solution to automatic emotion recognition for the Audio/Visual Emotion Challenge (AVEC) 2018. The objective is to find a robust way to detect emotions using more reliable emotion annotations in the valence and arousal dimensions. The two main contributions of this paper are: 1) the proposal of a new approach capable of generating more dependable emotional ratings for both arousal and valence from multiple annotators by extracting consistent annotation features; 2) the exploration of the valence and arousal distribution using outlier detection methods, which shows a specific oblique elliptic shape. With the learned distribution, we are able to detect the prediction outliers based on their local density deviations and correct them towards the learned distribution. The proposed method performance is evaluated on the RECOLA database containing audio, video and physiological recordings. Our results show that a moving average filter is sufficient to remove the incidental errors in annotations. The unsupervised dimensionality reduction approaches could be used to determine a gold standard annotations from multiple annotations. Compared with the baseline model of AVEC 2018, our approach improved the arousal and valence prediction of concordance correlation coefficient significantly to respectively 0.821 and 0.589.

AVEC'18, October 22, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5983-2/18/10…$15.00
https://doi.org/10.1145/3266302.3266307

CCS CONCEPTS

• **Human-centered computing** → *User models*;

KEYWORDS

emotion recognition, arousal valence distribution, feature agglomeration, outlier detection, multimodal fusion

ACM Reference Format:
Chen Wang, Phil Lopes, Thierry Pun, and Guillaume Chanel. 2018. Towards a Better Gold Standard: Denoising and Modelling Continuous Emotion Annotations Based on Feature Agglomeration and Outlier Regularisation. In *2018 Audio/Visual Emotion Challenge and Workshop (AVEC'18), October 22, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3266302.3266307

1 INTRODUCTION

Human emotions are often expressed through various modalities such as visual information (e.g. facial expressions and gestures), audio cues (e.g. tone, pitch and speed) and bodily responses (e.g. heart rate and skin conductance). The ability to recognise human emotions can enhance human computer interactions, allowing systems to use this information for personalisation and adaption towards users' affective states. Even though numerous researchers have worked on this topic through the application of discrete or dimensional emotion models, automatic emotion recognition is still a challenging task.

The Gold-Standard Emotion Sub-Challenge (GES) is a competition organised for the Audio/Visual Emotion Challenge (AVEC) workshop. Its objective is to increase the authenticity of emotional annotations and to improve the performance of automatic emotion recognition in the valence/arousal space using multimodal and data-driven approaches.

Detecting emotions from different multimodal affective expressions is a challenging task. Each modality has different time windows to reflect emotions. For example, facial expressions (visual modality) change faster than heart rate (physiological modality). Emotions are also highly dependent on the individual. Some people cry out feeling sad, while others keep a neutral expression in an attempt to hide their true feelings [23], adding to the difficulty of recognising emotions. Research on previous AVEC challenges have explored various methods: extracting multimodal features [6,

35], the fusion of multiple modalities [8, 34] and numerous deep learning architectures [33, 43] for recognising emotions.

Aside from modality and individual differences, the difficulties also come from emotion annotation. Emotion labels are obtained using two diverging methods: self-report or through expert (third-party) annotations. To obtain reliable annotations, different annotators are tasked in annotating the same set of emotional expressions. However, this presents problems of consistency, as reported values often diverge among the different annotators. They have different pre-existing emotional knowledge and reaction times, which can lead to time dissonance amongst annotations as shown in Figure 1.

Psychologists proposed different theoretical models for discrete emotions on the valence-arousal distribution, such as in [24, 25, 37]. These distributions could be useful for improving emotion prediction and classification. In practice, with manually assigned emotion labels, valence and arousal are observed to be correlated [10, 32]. However, for continuous emotions, the distribution of valence and arousal does not always match the theoretical models, which may be caused by labelling noise, stimulus bias and so on. Due to the noise in practice, the distribution models may not be applicable to emotion annotations or predictions directly.

Only a few studies have explored the pre-processing and post-processing of the distribution of arousal/valence annotations [18, 41, 28]. There are a few papers working on rating annotators or maximising the mutual information of multiple annotations [5, 19, 18]. However, so far there is no study which compares the performance of several denoising and dimensionality reduction methods.

The contributions of this work are mainly two-fold:

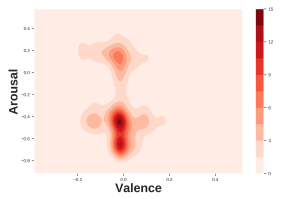
- First, we propose a new approach to obtain reliable emotion annotations from multiple annotators. To accomplish this task we started by applying a moving average to smooth each individual annotation, and subsequently extract their properties using several dimensionality reduction methods.
- Secondly, we study the distribution of arousal and valence in an unsupervised way and use it to regularise emotion predictions from multiple modalities. To the best of our knowledge, using the dimensional emotion distribution to support data-driven methods has not yet been explored. Outlier detection methods, such as the Local Outlier Factor (LOF), are applied to learn the local density deviation of a given data point with respect to its neighbours [7]. Once the local outlier factor has been learned, we use it as a regulariser to improve the prediction accuracy.

2 RELATED WORK

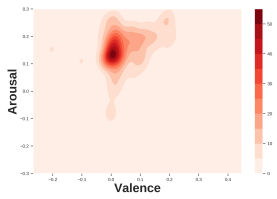
Emotion recognition is a multidisciplinary topic that has been explored by many researchers including psychologists, computer scientists and neuroscientists. In this section, we present the related work on modelling emotion distributions, annotation correction and multimodal recognition aspect respectively.

2.1 Distribution of Valence and Arousal

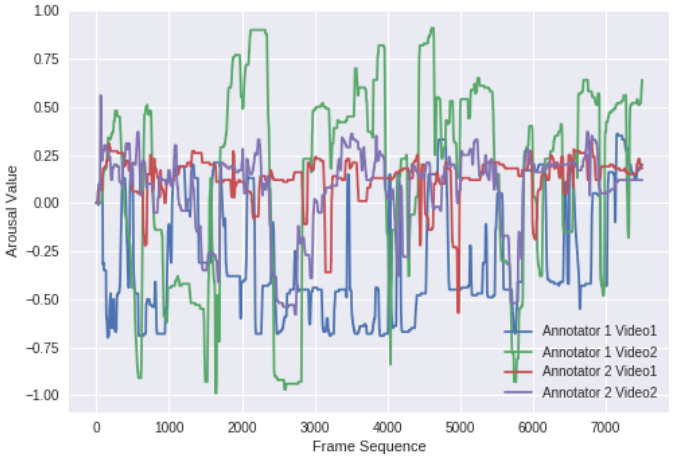
Dimensional models of emotion have attracted significant attention, providing a methodology for annotating different "degrees" of emotional intensity in a continuous fashion. In 1980, Russell [37]



(a) Annotation density distribution from Annotator 1 on Video 1 (darker red color indicates more points locate in that area)



(b) Annotation density distribution of from Annotator 2 on Video 1



(c) Annotation values from Annotator 1 and 2 on Video 1 and 2

Figure 1: Emotion annotation distribution on valence and arousal from RECOLA dataset [36]

formulated that emotions have a circumplex distribution in arousal and valence as shown in Figure 2. The theory assumes that different emotions are uniformly dispersed on a two-dimensional circular space (i.e. arousal and valence). This theory is widely accepted by many researchers [15, 22, 20]. Later Barrett et al [14] presented a bipolar distribution in valence and arousal replacing the circumplex model. In the bipolar distribution, as shown in Figure 1(a), emotions present an ascending bipolar continuum of valence, that varies from negative to most positive [25]. This bipolar shape has been applied in [32, 26] as prior knowledge for choosing the emotional stimuli. However, psychological studies showed that self-reported degrees of happiness and sadness do not correlate [42] suggesting that the bipolar model may be oversimplified. That leads to research such as the independent model [24] where positive and negative valence are independent dimensions and do not share a common axis (shown in Figure 2 (B)). While some researchers assume that valence and arousal are independent, it has been shown that there is actually a strong dependency between them. For instance [10] has found that the distribution of annotations in the valence/arousal space is U-shaped, as shown in Figure 2(C).

In practice, emotion labelling obtained from humans is slightly different from the aforementioned representations. This may be caused by the noise in the handcrafted emotion annotations. In [20], [44] and [1], the labelled data presents a similar oval shape, instead

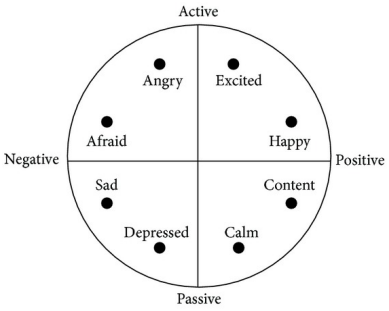


Figure 2: Circumplex Dimensional Model[37]

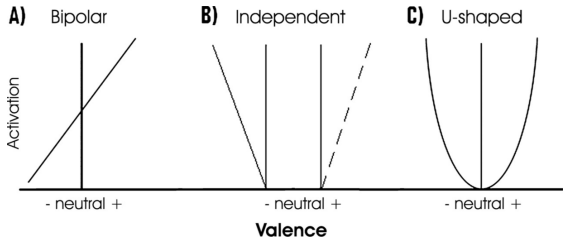


Figure 3: Valence-arousal Dimensional Model[25]

of a strict bipolar distribution or a U shape. Also, the research so far was on discrete emotion labels instead of continuous annotations. This work will attempt to model these distributions using real and continuous arousal and valence ratings for the construction of statistical and data-driven models.

2.2 Emotion Annotation Correction

The downside of emotion annotations is that they can often suffer from noise, due to the different emotional perception and reaction time of the different annotators. This can lead to temporal dissonance between the different annotations. Furthermore, annotation errors can also occur such as miss-clicks or misapprehensions, which can increase inaccuracy. Thus, a processing algorithm exhibiting robustness and fault tolerance to manual annotations is needed. Mariooryad and Busso [28] proposed to compensate annotation dissonance by using a time-shift, which maximises the mutual information between the emotion expressions and the annotations. In [41], three different filters (moving average, Savitzky-Golay, and a median filter) were applied to smooth the annotation data, where the moving average filter was suggested as a promising method for increasing performance. Annotation fusion by [31], [5] applied time wrapping method and additional comparative rank-based information, which obtained better recognition results. While in [19], it took the uncertainty of human emotion perception into account based on the inter-rater disagreement level. [18] proposed an expectation maximisation approach to model specific distortions from each annotator such as perception bias and delay. It is important to note that this particular research topic is still young, and more work on this topic is still needed.

2.3 Multimodal Emotion Recognition

Previous research has shown that different deep learning architectures such as the auto-encoder, Convolutional Neural Network (CNN) and Deep Belief Network (DBN), can generate robust features effectively from a wide range of modalities, capable of revealing complex hidden cues of emotion. A large portion of the research focuses on basic emotion classification (e.g. [34] and [17]). Recently, databases such as SEMAINE [29] and RECOLA [36] with time-continuous emotion ratings have shifted the methods from classification to regression to predict continuous emotion in several emotion dimensions.

Models for dimensional emotion recognition can be classified into two categories. The first class is aimed at feature extraction whilst the second one specialises more in emotion recognition with multiple modalities. In the first category, different levels of features are derived from audio, visual and physiological signals. Previous work has extensively explored audio features from an acoustic, functional and linguistic perspective in [38]. The interlocutor influence has been taken into consideration, while extracting the audio features by [8]. Currently, deep convolutional neural networks (CNN) are the state-of-the-art models for extracting visual features ([6, 40, 21]). A deep autoencoder network has been proposed by Ngiam et al [30] to extract features from both audio and video modalities to predict emotions and has shown to be promising. Although for the physiological modality, this has rarely been explored. Currently, features from physiological signals are mainly low level such as the heart rate variability [2].

To predict the dimension emotions, the second category of models contains non-temporal and temporal models. The non-temporal models usually require contextual features while temporal models emphasise the dynamic information in the model directly. Long Short Term Memory (LSTM) models are currently widely used for temporal models, where several topologies are explored [16, 6, 8]. For the emotion prediction task, it is necessary to determine the appropriate length of temporal windows, which can vary based on the modality [43]. For example, according to [17, 16, 35], audio signals change faster over time than video signals and physiological signals. To take full advantages of different modalities, fusion techniques can be applied at feature, decision or modality level. According to [17, 8, 43] and [30], the multimodal recognition methods outperform unimodal methods significantly.

It is apparent from literature that deep learning methods have been extensively used for the task of multimodal emotion recognition. In contract, the approaches of annotation correction and dimensional emotion distribution, still needs more exploration.

3 PROPOSED METHOD

In this section, we introduce our methods for annotation correction and dimensional emotion distribution learning. Figure 4 showcases the pipeline of our proposed system: emotion labels are learned from annotators' annotations as well as multimodal information. With annotation correction methods, we get a set of intermediate arousal and valence labels for the training dataset. Emotion recognition will be trained with multimodal features and the intermediate labels. The distribution of arousal and valence is learned from the same set of labels as well. For the validation dataset, arousal and valence

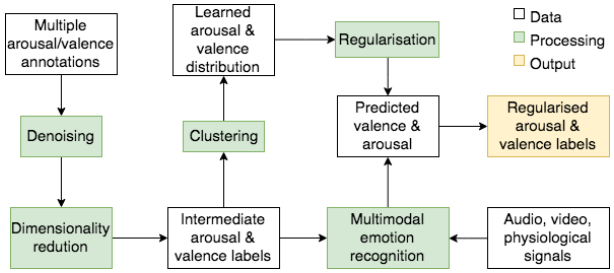


Figure 4: Schematic diagram of proposed method

values are predicted from multimodal features and then regularised by the learned arousal and valence distribution. At the same time, the distribution model will be updated with the new set of labels. In this way, we get the final emotion labels represented by the yellow box in Figure 4.

3.1 Annotation Correction

Our label correction methodology analyses annotations from two different perspectives: the individual perspective and the group perspective. The individual perspective consists of a denoising methodology for each individual annotation (i.e. filtering). The group perspective combines the multiple annotations in a single output label. It contains dimensionality reduction methods for finding consistent emotion labels.

At the individual level, it is inevitable that each annotation from a given annotator contains noise owing to the unfamiliarity with the annotation tool, the arousal/valence concept, or simply mistaken mouse clicking [35]. It is reasonable to hypothesise that people do not get from one emotional state to another instantly [41], but there is a gradual passage that takes them from one state to the next. Therefore, when there are very sudden changes in the annotation values, they are more likely to be noise. For the RECOLA dataset, as shown in Figure 1(c), each annotation curve presents some sharp peaks superimposed on a smooth changing trend. The aforementioned noise can be removed by applying sliding window filters on the individual annotation data. The window type, size and sliding step all influence the denoising performance and information loss. We implemented a 1D convolution with several window types (flat, median and Hanning) and a window size ranging from 2 to 10 seconds. The window size was chosen based on previous work ([35, 41]).

At the group level, annotators have their unique way of perceiving and reporting emotions. Analysing Figure 1(a) and 1(b), we can clearly see that annotations have very different distributions, even on the same video. Combined with Figure 1(c), it reflects the inconsistencies between the different annotations. Even after applying the individual level process, these inconsistencies between annotators are still apparent, as demonstrated in Table 1, which shows the valence-arousal 2D Pearson correlation coefficients. Thus we propose to apply unsupervised dimensionality reduction methods on annotations from all annotators to get one set of reliable representation of arousal and valence. To be more concrete, we tested sparse principal component analysis and feature agglomeration methods.

Sparse Principal Component Analysis (SPCA) [47] extracts the sparse components that best reconstruct the multivariate dataset. SPCA is an extension of classic principal component analysis (PCA) for reducing data dimensionality by adding sparsity constraints on the inputs. One big advantage of SPCA is that it allows to choose the leading principal axis, which means the obtained low-dimension data can be interpretable. Also compared with ordinary PCA, SPCA can find linear combinations that contain fewer input variables. For the implementation, a minimal reconstruction error approach is applied following [47].

Assuming that all annotations on the same video contain consistent emotion features, feature agglomeration ([39]) was also tested. Feature agglomeration is a hierarchical clustering method that merges similar features to find a low-dimensional representation of the original data. It starts from the 'bottom' where each observation is assigned to its own cluster. Then pairs of clusters are merged as one and moves up the hierarchy. For the merge strategy, we used recursive Ward's method [46] to merge the pair of feature clusters that minimally increases within-cluster variance. The Euclidean distance $d_{(ij)}$ between cluster C_i and C_j is updated as:

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|$$

where d_{ik} and d_{jk} are the pairwise distances between clusters C_i, C_j, and C_k. Parameters α_i, α_j, β and γ depend on cluster sizes. The feature agglomeration methods tend to have uneven cluster sizes without constrains. These parameters are used as cluster connectivity constrains [46] and are updated at each step when a pair of clusters is merged.

3.2 Outlier regularisation on predicted labels

As mentioned in Section 2.1, arousal and valence are highly, though not completely, correlated. We calculated the mean Pearson correlation coefficient between arousal and valence on the training and validation subsets of RECOLA. Using the processed annotations obtained through the methodology presented in 3.1, we observed a correlation for arousal and valence of 0.37 and 0.41, respectively. We also observed the distribution density of the two dimensions on the whole training dataset, shown in Figure 8. The density observed in Figure 8 showcases a similar distribution as in [20], [44] and [1]. There is a whole research branch on learning data distributions. For this work, the joint arousal/valence distribution is learned in order to improve emotion prediction accuracy. To learn this distribution, we propose the Local Outlier Factor (LOF) method.

This particular algorithm assumes that the density around an outlier object is significantly different from the density around its neighbours. Thus the k-NN algorithm [7] is used to calculate the local density between neighbouring points. As shown in Figure 5, point D has a lower local density than point A. If the red circle is the learned decision function for the dimensional emotion distribution, the distance between point A and B is the maximum distance. Thus point D is out of the reachablity distance and considered as an outlier. Several other outlier detection methods were also tested such as: the isolation forest[27], one-class support vector machine (SVM) [9] and robust covariance [45].

To improve the emotion prediction accuracy and correct prediction outliers, we can check each pair of predicted valence and arousal with the learned distribution. If one pair of predictions are

Table 1: 2D Pearson Correlation Coefficients Between Annotators on Arousal and Valence

Pearson CC	Annotator1	Annotator2	Annotator3	Annotator4	Annotator5	Annotator6
Annotator1	1	-0.0719444	0.414125	0.491069	0.0446386	-0.127891
Annotator2	-0.0719444	1	0.212089	-0.195357	0.212211	0.334771
Annotator3	0.414125	0.212089	1	0.190809	0.263295	0.192494
Annotator4	0.491069	-0.195357	0.190809	1	-0.217903	-0.363663
Annotator5	0.0446386	0.212211	0.263295	-0.217903	1	0.259848
Annotator6	-0.127891	0.334771	0.192494	-0.363663	0.259848	1

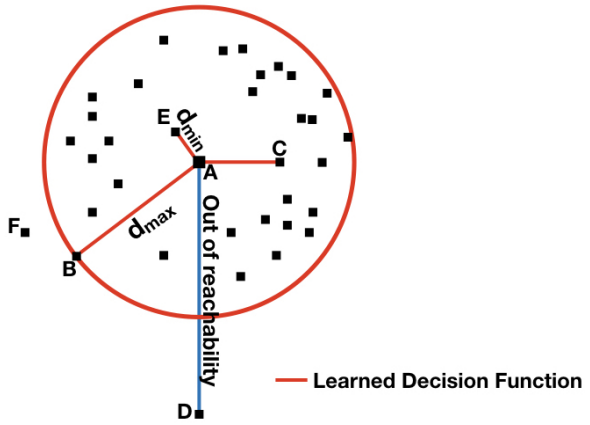


Figure 5: Local Outlier Factor: Each point is compared with its local neighbours instead of the global data distribution

detected as an outlier, we first check the prediction points from their connective time frames. If neither of these points are outliers, we take the average of them and regularise the outlier by taking the average value. Otherwise, we find the nearest neighbour and correct the outlier towards that direction, using half of the Euclidean distance between the two points. This regularisation is based on the assumption that emotions do not change abruptly [41].

Since the main goal of this study is not to propose a new method to map the multimodal feature space to the targets, we used the baseline emotion recognition system from AVEC 2018 [13] to get the predicted dimensional emotion values and to evaluate the proposed methods.

4 EXPERIMENTS AND RESULTS

The main purpose of this study is to examine the feasibility of learning and regularising annotations in order to improve emotion recognition performance. As illustrated in Figure 4, we start by validating our proposed methods on the annotation data from RECOLA [36]. Subsequently, we investigated the emotion recognition performance using the baseline model from AVEC 2018 on the pre-processed annotations.

4.1 Dataset and Baseline Model

The RECOLA dataset from the AVEC 2018 challenge [36] is used for this study. It includes audio, video and physiological recordings (electro-cardiogram and electro-dermal activity data) from 27 French-speaking subjects aged between 18 and 25 years old. We

Window	2s	6s	10s
Arousal	0.97	0.86	0.78
Valence	0.98	0.94	0.88

Table 2: Average R^2 result of different window sizes of the training set

used the subset offered by the GES subchallenge, which contains 9 subjects for the training dataset and another 9 subjects for validation. The remaining 9 subjects are used as the testing set for the challenge. The results reported below are all evaluated from the validation set. The duration of each recording is 5 minutes. Arousal and valence are annotated by six annotators every 400 ms (each video frame) and the annotation values are scaled between [-1, +1]. The baseline emotion recognition model [13] provided by AVEC 2018 extracted features with several open-source tools such as openSMILE [11] and openFace [4]. The audio baseline features contains Bag of Audio Words (BoAW) [11], DeepSpectrum[3] and Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features [12]. Physiological features includes mean heart rate and Heart Rate Variablity (HRV), Skin Conductance Level (SCL) and Skin Conductance Response (SCR) while feature extracted from videos include: appearance, action units, Box of Visual Words (BoVW) and geometric features. To evaluate the algorithm performance, the concordance correlation coefficient (CCC) is used, which is defined as:

$$CCC = 2 \times \frac{COVAR[X,Y]}{(VAR[X]} + (E[X] - E[Y])^2) \quad (1)$$

where X denotes the prediction values and Y the gold standard values.

4.2 Results of Annotation Correction

The information loss is evaluated when the sliding window is applied. For each annotation, the data is filtered with flat (moving average), median and Hanning windows with a small, medium, and large window size, which contains 50, 150, and 250 samples respectively. The window slides forward with a constant single frame (40 ms). Figure 6 presents an example of applying the filtering technique to the annotation data of a video from the RECOLA dataset. The result of this interpolation was compared with the original rating by computing CCC. Pairwise t-test is applied to check whether the difference of performance is significant.

Results (2) show that with the size of the sliding window increasing, the information loss increases faster on arousal than on valence,

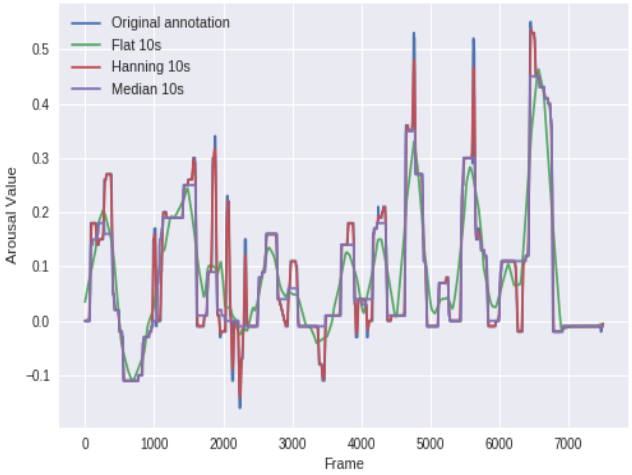


Figure 6: Filtering with three types of sliding windows: flat, median and Hanning

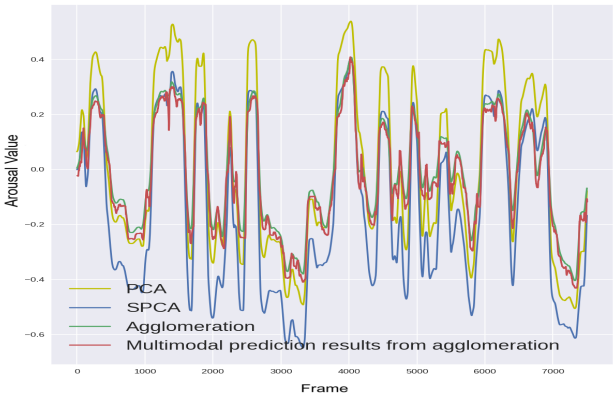


Figure 7: Results of dimensionality reduction

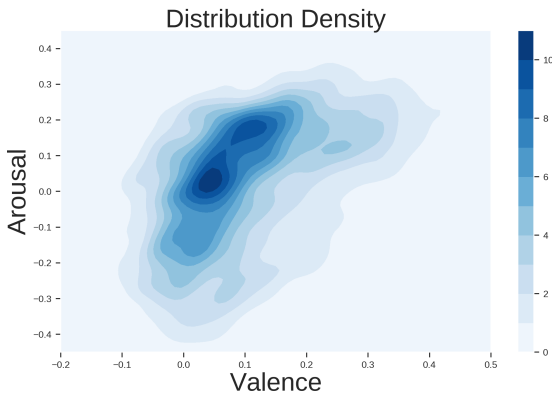


Figure 8: Density Distribution of Arousal and Valence

which indicates that the emotional valence changes less rapidly than arousal over time [35]. The filter could remove high-frequency fluctuations noticeably in the annotations. Large window size resulted in distinct distortion in the outputted annotation data compared with the original annotation data. Comparing the flat window to the Hanning window, the later conserves more high-frequency fluctuations, while the former leads to a smoother annotation curve. Each annotation is subsequently normalised through z-score.

We evaluated the window influence on emotion recognition with the baseline model.The performance (CCC) improved when the window size increased for both Hanning and flat windows, which confirms the findings in [41]. While for median window, the performs drops with the largest window size. Both median and flat window outperform the Hanning window with an identical window size. That indicates that the high frequency information kept by the Hanning window is more likely to be noise. The median filter with window size equals to 6 seconds achieved the best CCC performance.

It can be seen in Figure 7 that the annotation learned from dimensionality reduction methods have analogous trends with different variance. Comparatively to the baseline method (arousal CCC = 0.775 and valence CCC = 0.57), we can observe that feature agglomeration achieved a higher CCC, 0.816 for arousal and 0.583 for valence. While SPCA has a worse performance with 0.622 on arousal and 0.505 on valence. The SPCA method maximises the variance among annotations, and feature agglomeration merges the two nearest clusters, which minimises the variance within the merged new cluster. That indicates that the consistency among annotators contains more reliable emotion information both for valence and arousal.

4.3 Distribution of Arousal and Valence

In order to learn the joint distribution of arousal and valence, we compared several widely applied outlier detection methods. In this analysis both the supervised and unsupervised methodologies are included: local outlier factor (LOF), one-class SVM, isolation forest, and robust covariance. As shown in Table 3 and Figure 10, local outlier factor outperforms the other methods significantly. Although the difference on the false-positive ratio is not prominent with an isolation forest, the LOF detected much more outliers when tested on the validation set with 8000 points. LOF determines outliers based on the local density deviation, which is calculated by k nearest neighbours. Usually k is selected between 20 to 30 [7]. We set k to 50 with the assumption that emotion does not change much within 2 seconds (50 frames). When using the learned decision function as a regularizer for the predicted values, it significantly improves the unimodal (audio, video and physiological modality) prediction by 0.061 (T=3.51, p=.017). The improvement on arousal is shown in Figure 9. While for the multimodal prediction, the improvement is minor. The multimodal fusion significantly improves the prediction accuracy (T=6.79, $p < 0.01$) compared with a unique modality. Consequently, the inaccurate predictions of the multimodal fusion are mostly located within the learned arousal/valence cluster. Consequently they are not recognised as outliers and not regularised.

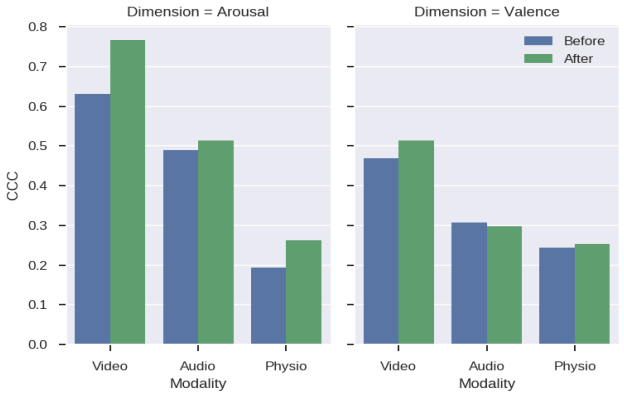


Figure 9: CCC of Arousal and Valence: Unimodal prediction before and after arousal-valence regularisation.

Table 3: False Positive Ratio of Outlier Detection (Bold:best performance; *: p<0.02)

LOF	Random Forest	One-class SVM	Robust Covariance
5.6	5.7	10.1*	10.7*

5 CONCLUSIONS AND FUTURE WORK

This paper provides a new approach to the AVEC 2018 GES challenge [13] by extracting robust annotations from multiple manual emotion labels. This approach consists of sliding window annotation denoising, feature agglomeration on multiple annotations and local outlier factor regularisation of predicted arousal-valence values. Our approach outperforms the baseline method introduced by the AVEC 2018 challenge. The concordance correlation coefficient (CCC) on valence and arousal achieved by our method are 0.589 and 0.821 respectively, while the baseline results are 0.57 and 0.775. Our primary goal is to enhance emotion recognition performance by improving the robustness of annotation data before proceeding into emotion recognition. Empirical results showed that our proposed methods for treating individual and multiple annotations improved emotion recognition results to 0.821 and 0.589 on arousal and valence respectively. Learning the arousal-valence distribution and applying an outlier detection method as a regularisor significantly improved the performance from unimodal emotion prediction. However, it is important to note that this work could still be improved in the following ways. This work was tested exclusively on the RECOLA database with six annotators. The proposed method may be overfitted on this particular dataset. Further experimentation is necessary to improve the robustness of the suggested approach and validate its performance on different databases. Combining the time-series of dynamic annotation features and the valence-arousal spatial space is an interesting solution for emotion recognition. On top of that, evaluating and judging the quality of a gold standard continues to be an open question. Even if results improve through the proposed methods, it does not guarantee that the processed annotation will be "authentic information".

In the future, we would like to explore the multiple continuous annotations with temporary models and combine with the 2D spatial distribution in the valence/arousal space proposed in this paper. In this way, the deep learning architectures trained on video recognition tasks could be transferred on annotation processing, which may generate useful features for recognising emotions.

ACKNOWLEDGMENTS

This work is supported by the Swiss National Science Foundation under Grant No.: 2000221E-164326.

REFERENCES

[1] Soraia M Alarcão and Manuel J Fonseca. 2017. Identifying emotions in images from valence and arousal ratings. *Multimedia Tools and Applications*, 1–23.
[2] Mohammadreza Amirian, Markus Kächele, Patrick Thiam, Viktor Kessler, and Friedhelm Schwenker. 2016. Continuous multimodal human affect estimation using echo state networks. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 67–74.
[3] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. 2017. Snore sound classification using image-based deep spectrum features. In *Proceedings INTERSPEECH*, 3512–3516.
[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.
[5] Brandon Booth, Karel Mundnich, and Shrikanth S Narayanan. 2018. A novel method for human bias correction of continuous-time annotations.
[6] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S Huang. 2016. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 97–104.
[7] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. Lof: identifying density-based local outliers. In *ACM sigmod record* number 2. Vol. 29. ACM, 93–104.
[8] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal multitask learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 19–26.
[9] Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. 2001. One-class svm for learning in image retrieval. In *Image Processing, 2001. Proceedings. 2001 International Conference on*. Vol. 1. IEEE, 34–37.
[10] William A Cunningham, Carol L Raye, and Marcia K Johnson. 2004. Implicit and explicit evaluation: fmri correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of cognitive neuroscience*, 16, 10, 1717–1729.
[11] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.
[12] Florian Eyben et al. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7, 2, 190–202.
[13] Fabien Ringeval and Björn Schuller and Michel Valstar and Roddy Cowie and Heysem Kaya and Maximilian Schmitt and Shahin Amiriparian and Nicholas Cummins and Denis Lalanne and Adrien Michaud and Elvan Çiftçi and Hüseyin Güleç and Albert Ali Salah and Maja Pantic. 2018. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. In *Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge, AVEC'18, co-located with the 26th ACM International Conference on Multimedia, MM 2018*. Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic, (Eds.) ACM, Seoul, Korea, (Oct. 2018).
[14] Lisa Feldman Barrett and James A Russell. 1998. Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology*, 74, 4, 967.
[15] Andrew J Gerber et al. 2008. An affective circumplex model of neural systems subserving valence, arousal, and cognitive overlay during the appraisal of emotional faces. *Neuropsychologia*, 46, 8, 2129–2139.
[16] Hatice Gunes and Maja Pantic. 2010. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)*, 1, 1, 68–99.
[17] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: current trends and future directions. *Image and Vision Computing*, 31, 2, 120–136.

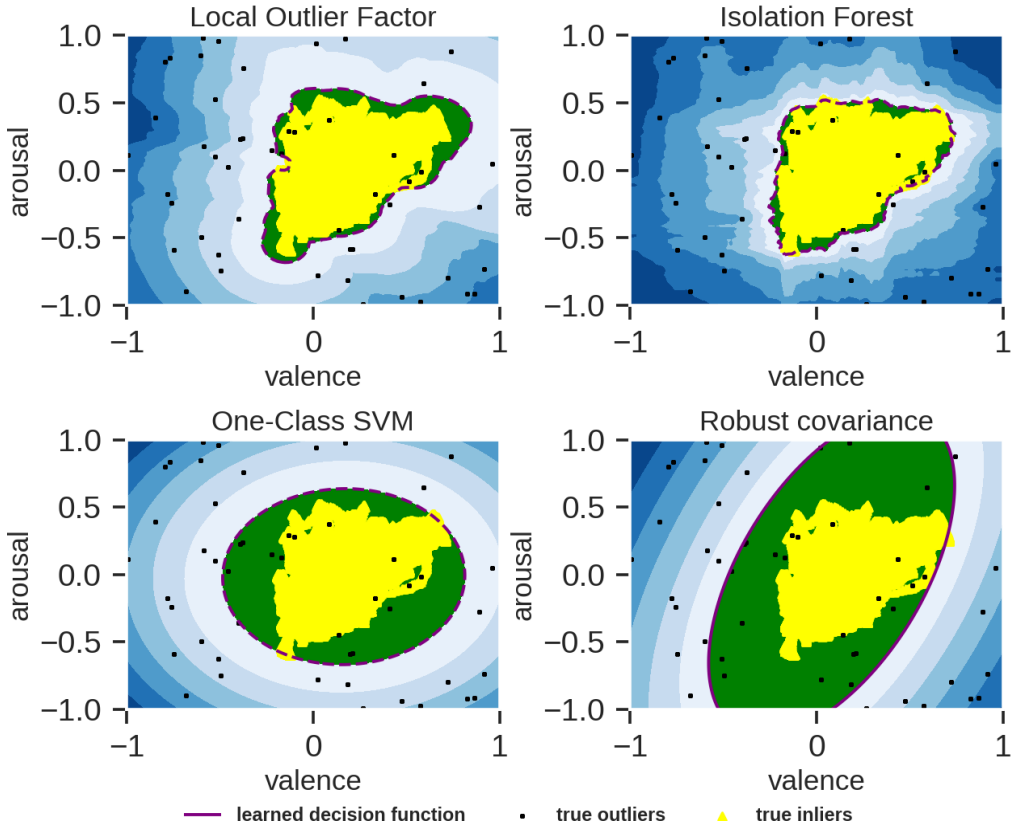


Figure 10: Outlier Detection

[18] Rahul Gupta, Kartik Audhkhasi, Zach Jacokes, Agata Rozga, and Shrikanth Narayanan. 2018. Modeling multiple time series annotations as noisy distortions of the ground truth: an expectation-maximization approach. *IEEE transactions on affective computing*, 9, 1, 76.

[19] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From hard to soft: towards more human-like emotion recognition by modelling the perception uncertainty. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 890–897.

[20] Suwicha Jirayucharoensak, Setha Pan-Ngum, and Pasin Israsena. 2014. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014.

[21] Markus Kächele, Patrick Thiam, Günther Palm, Friedhelm Schwenker, and Martin Schels. 2015. Ensemble methods for continuous affect recognition: multimodality, temporality, and challenges. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 9–16.

[22] JungHyun Kim, Seungjae Lee, SungMin Kim, and Won Young Yoo. 2011. Music mood classification model based on arousal-valence values. In *Advanced Communication Technology (ICACT), 2011 13th International Conference on*. IEEE, 292–295.

[23] Ann M Kring, David A Smith, and John M Neale. 1994. Individual differences in dispositional expressiveness: development and validation of the emotional expressivity scale. *Journal of personality and social psychology*, 66, 5, 934.

[24] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. 1993. Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30, 3, 261–273.

[25] Penelope A Lewis, HD Critchley, P Rotshtein, and RJ Dolan. 2006. Neural correlates of processing valence and arousal in affective words. *Cerebral cortex*, 17, 3, 742–748.

[26] Elizabeth A Linnenbrink. 2007. The role of affect in student learning: a multidimensional approach to considering the interaction of affect, motivation, and engagement. In *Emotion in education*. Elsevier, 107–124.

[27] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 413–422.

[28] Soroosh Mariooryad and Carlos Busso. 2015. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6, 2, 97–108.

[29] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3, 1, 5–17.

[30] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689–696.

[31] Mihalis A Nicolaou, Vladimir Pavlovic, and Maja Pantic. 2014. Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. *IEEE transactions on pattern analysis and machine intelligence*, 36, 7, 1299–1311.

[32] Reinhard Pekrun, Thomas Goetz, Anne C Frenzel, Petra Barchfeld, and Raymond P Perry. 2011. Measuring emotions in studentsfffdfffdfffd learning and performance: the achievement emotions questionnaire (aeq). *Contemporary educational psychology*, 36, 1, 36–48.

[33] Filip Povolny, Pavel Matffffdfffdejka, Michal Hradis, Anna Popková, Lubomír Otrusina, Pavel Smrz, Ian Wood, Cecile Robin, and Lori Lamel. 2016. Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 75–82.

[34] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–9.

[35] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66, 22–30.

[36] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 1–8.

[37] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39, 6, 1161.

[38] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Communication*, 53, 9-10, 1062–1087.

[39] Samir Shah and Arun Ross. 2006. Generating synthetic irises by feature agglomeration. In *Image Processing, 2006 IEEE International Conference on*. IEEE, 317–320.

[40] Bo Sun, Siming Cao, Liandong Li, Jun He, and Lejun Yu. 2016. Exploring multimodal visual features for continuous affect recognition. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 83–88.

[41] Nattapong Thammasan, Ken-ichi Fukui, and Masayuki Numao. 2016. An investigation of annotation smoothing for eeg-based continuous music-emotion recognition. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*. IEEE, 003323–003328.

[42] Silvan Solomon Tomkins and Carroll Ellis Izard. 1965. Affect, cognition, and personality: empirical studies.

[43] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11, 8, 1301–1309.

[44] Andero Uusberg, Helen Uibo, Riti Tiimus, Helena Sarapuu, Kairi Kreegipuu, and Juri Allik. 2014. Approach-avoidance activation without anterior asymmetry. *Frontiers in Psychology*, 5, 192. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.00192. https://www.frontiersin.org/article/10.3389/fpsyg.2014.00192.

[45] Ratko G. Veprek, Sebastian Steiger, and Bernd Witzigmann. 2007. Ellipticity and the spurious solution problem ofkfffdfffdfffdpenvelope equations. *Physical Review B*, 76, 16. DOI: 10.1103/physrevb.76.165320.

[46] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58, 301, 236–244.

[47] Hui Zou, Trevor Hastie, and Robert Tibshirani. 2006. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15, 2, 265–286.