An Active Feature Transformation Method For Attitude Recognition of Video Bloggers

Fasih Haider¹, Fahim A.Salim², Owen Conlan² and Saturnino Luz³

¹IPHSI, University of Edinburgh, UK ²ADAPT Centre, Trinity College Dublin, Ireland ³{Fasih.Haider, S.Luz}@ed.ac.uk, {salimf,owconlan}@scss.tcd.ie

Abstract
Video blogging is a form of unidirectional communication where a video blogger expresses his/her opinion about different issues. The success of a video blog is measured using metrics like the number of views and comments by online viewers. Researchers have highlighted the importance of non-verbal behaviours (e.g. attitudes) in the context of video blogging and showed that it correlates with the level of attention (number of views) gained by a video blog. Therefore, an automatic attitude recognition system can help potential video bloggers to train their attitudes. It can also be useful in developing video blogs summarization and searching tools. This study proposes a novel Active Feature Transformation (AFT) method for automatic recognition of attitudes (a form of non-verbal behaviour) in video blogs. The proposed method transforms the Mel-frequency Cepstral Coefficient (MFCC) features for the classification task. The Principal Component Analysis (PCA) transformation is also used for comparison. Our results show that AFT outperforms PCA in terms of accuracy and dimensionality reduction for attitude recognition using linear discrimination analysis, 1-nearest neighbour and decision tree classifiers.

Index Terms: Feature Engineering, Feature Transformation, Feature Extraction, Attitude Recognition, Video Bloggers

1. Introduction
Video blogs (Vlogs) are a popular form of unidirectional communication through social media where the vlogger (video blogger) does not receive feedback from the viewers in real time, but viewers can provide their feedback later in the form of textual comments. Studies conducted on video blogs such as [1] concluded that the non-verbal behavior of the vlogger influences the level of attention gained by a video. Therefore, it is useful to automatically analyze non-verbal behavior in vlogs to gauge the vlogger’s behavior and provide feedback so that they can improve their vlogs. In addition, automatic recognition of attitudes could also help in developing vlogs summarization and searching tools. In this study, we investigate attitude (states that may permeate strong emotions [2]) recognition.

In the discipline of affective computing, researchers have proposed many techniques to detect emotional states in different contexts ranging from human-human to human-machine communication [3, 4, 5]. However, analysis of vlogs has not been explored extensively in the literature. In one study, the facial expression, acoustic (speaking activity and prosodic features) and the multimodal information is used to predict the personality traits in vlogs using regression analysis [6]. A perceptual and acoustic analysis is performed for 12 different attitudes expressed by Portuguese speakers [7]. The results show that the audio-visual information provides a better perception of attitudes than any single modality. An analysis of speaking time, F0 energy, voice rate, speech turn along with head motions, looking time and proximity to camera in terms of Pearson’s correlation (between non-verbal cues and the median number of log views) by Biel et al. [8] showed that the audio-visual cues are significantly correlated with the median number of log views.

An automatic attitude detection system for multimodal dialogue systems is proposed in [9] which used acoustic features. In the analysis on a subset of the data in [9], Madzlan et al. [10] analyzed the acoustic and high-level visual features (facial landmarks) to train a classifier to detect the attitude automatically. In it, the authors propose a three-class problem grouping the attitudes in the following three classes: positive, negative and neutral attitudes [10]. They defined friendliness attitude as neutral. Amusement and enthusiasm as positive attitudes, and ‘frustration and impatience’ as negative attitudes. The results show that the acoustic features (63.63%) provide better results than the visual features (50.6%). However, they did not perform fusion of features. In a different study [11], authors analyzed prosodic features of vlogger and found that these features (F0, voice quality and intensity) are correlated with a vlogger attitude, while in [12] they analyzed audio-visual features of vloggers for their attitude recognition. In all of the above studies, authors extracted the acoustic features using statistical functions (e.g. mean, standard deviation, maximum, minimum of prosodic and voice quality features) over a speech segment level (where the speech segment is a speech utterance). Ergodic Hidden Markov Models (HMM) are also employed to generate a representation over a speech segment level [13] but there is no clear interpretation of HMM states for emotion recognition as for automatic speech recognition (sub phoneme) [14]. Haider et al. analysed the acoustic and visual features for attitude recognition using audio and visual (Fisher vector representation of dense histogram of gradient, dense histogram of flow and dense motion boundary histogram) features [15]. Mel-frequency cepstral coefficients have been widely used in applications related to speech processing like speech recognition, speaker recognition and spoken expressions recognition. MFCC features are the short-term power spectrum of the audio signal which is sensitive to noise and duration of the frame used for calculating the features. When MFCC features are used in speech recognition tasks, they are extracted over a frame level of fixed duration, is typically 10 ms to 40 ms for speech recognition. However, the current emotion/affect recognition approaches calculate a statistical response of MFCC features over a speech segment level, with typical duration of a few seconds. It is our contention that using a statistical response of MFCC at the speech segment level for affect/emotion recognition would lose significant information. This is because the speech signal properties vary considerably more over larger (segment-length) speech segments than small (100 ms) frames. To overcome this problem, we propose a novel feature extraction method which transforms MFCC features to a lesser dimension than a statistical function. This is
done using a frame size of 100 ms with 67% overlap, and then transforming the results to represent the speech segments using a machine learning method, namely, self-organizing mapping. To the best of our knowledge, there is no study which demonstrates the discrimination power of such MFCC features for attitude recognition of vloggers. The contributions of this study are therefore:

1. a demonstration of the discrimination power of MFCC features which are extracted over a speech segment level using statistical functions and their transformation by PCA for the recognition of six attitudes (Amusement (A), Enthusiasm (E), Friendliness (Fd), Frustration (Fr), Impatience (I) and Neutral (N)) of video bloggers, and
2. a novel Active Feature Transformation (AFT) method which can extract features at the speech segment level using a machine learning method (self-organizing maps) with reduced dimensionality. This method is evaluated for attitude recognition, as stated above.

The AFT can transform features which are extracted over speech segments of variable duration (variable duration of speech segments result in variable dimensions of features), where other feature extraction and selection methods require a fixed dimension data for transformation. To calculate a fixed dimension response for speech segments of variable duration, different approaches are used such as statistical-functional response and Fisher vector generation before deploying feature engineering and classification methods. In this study we are evaluating statistical-functional response over a speech segment level along with PCA against the AFT.

2. Active Feature Transformation Method

The Active Feature Transformation (AFT) method comprise of the following steps for features transformation:

1. First a speech segment \( (S_i) \) is divided into \( n \) frames \( \{ F_{k,S_i} \} \) of fixed duration (100 ms) with an overlap of 67% with the neighboring frame, where \( i = 1 : N \) and \( N \) represents the total number of speech segments (our case \( N = 613 \), as shown in Table 1), and \( k = 1 : n \), that is \( k \) varies from 1 to \( n \), the total number of frames in a speech segment \( (S_i) \). Hence \( F_{k,S_i} \) is the \( k \)-th frame of \( i \)-th speech segment and 228 MFCC features are extracted over a frame \( (F_{k,S_i}) \) instead of extracting them over speech segments of variable duration. The system architecture is depicted in Figure 2.

2. Clustering of frames: Instead of using statistical functionals of MFCC to reduce the dimensions for speech segments of variable duration, which results in loss of discriminating power, we used self organising maps (SOM) [16] for the clustering of frames into \( n \) clusters \( \{ C_1, C_2, ..., C_n \} \), as depicted in Figure 3. Here \( n \) represents the cluster size for SOM (in our case \( n = 5 \times 5 : 100 \)).

3. Generation of an Active Feature transformation \( (AFT_{S_i}) \) vector by calculating the number of frames in each cluster for each speech segment \( (S_i) \) as depicted in Figure 3.

4. As the number of frames are different for each speech segment (i.e. the duration of all speech segment is not constant), we normalise the feature vector by dividing it with the total number of frames present in each speech segment \( (\sum AFT_{S_i}) \) as set out in Equation 1.

\[
AFT_{S_i_{norm}} = \frac{AFT_{S_i}}{\sum AFT_{S_i}} \tag{1}
\]

3. Dataset

The video-blog dataset used in this study is the same used in [10] augmented with the annotation of a hundred segments with a neutral label [15]. In total, it contains the 613 audio-visual segments (for each subject the number of speech segments/instances is as follows: 34, 53, 54, 111, 46, 36, 93, 104, 34, 48) from around 250 different videos that are annotated for six different attitudes (Amusement-A, Enthusiasm-E, Friendliness-Fd, Frustration-Fr, Impatience-I and Neutral-N) as depicted in Table 1. The data annotation was performed by two annotators with an inter-coder agreement of 75% as reported in [17]. The data comes from 10 different native speakers of English. The duration of video clips is around 1-3 seconds. In this study, only the audio information is used.

Table 1: Number of instances (speech segments/speech utterances) for each attitude in the dataset

<table>
<thead>
<tr>
<th>Attitude</th>
<th>speech segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amusement</td>
<td>100</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>107</td>
</tr>
<tr>
<td>Friendliness</td>
<td>101</td>
</tr>
<tr>
<td>Frustration</td>
<td>103</td>
</tr>
<tr>
<td>Impatience</td>
<td>102</td>
</tr>
<tr>
<td>Neutral</td>
<td>100</td>
</tr>
</tbody>
</table>

4. Experimentation

This section describes the acoustic features extraction and classification methods.

4.1. Feature Extraction

We use the openSMILE [18] to extract the acoustic features using \textit{emobase.config} configuration file which has been widely used for emotion recognition [3]. In this study we considered MFCC features which are in total 228 (a subset of 988 features extracted using \textit{emobase.config}), and use three different features vectors for the following experiments:

1. Experiment one: 228 MFCC features extracted for each speech segment. OpenSMILE calculates an overall MFCC features response for speech segments with variable duration (1 - 3 seconds in this case) using statistical functionals such as mean, standard deviation, minimum, maximum, range values etc. The objective of calculating an overall response is to project the features on a fixed dimension (in this case it is 228) for machine learning methods (e.g. dimensionality reduction and classification).

2. Experiment two: Transformed version of 228 MFCC (extracted by openSMILE) for each speech segment using PCA.

3. Experiment three: Transformed version of 228 MFCC features, which are extracted for a frame of fixed duration using AFT as described in 2.
The classification is performed using three different methods namely Linear Discriminant Analysis (LDA), Nearest Neighbour (KNN with K=1) and Decision Trees (DT). These classifiers are employed in MATLAB\(^1\) using the statistics and machine learning toolbox in the 10-fold cross-validation setting. 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 [19]). KNN and DT are non-parametric, non-linear methods, included for comparison.

\(^1\)http://uk.mathworks.com/products/matlab/ (June 2017)
PCA and AFT, for datasets of dimensionality greater than 10 dimensions shows higher mean values for AFT than PCA, and this difference is statistically significant for LDA (\(p_{LDA} = 0.003\), \(mean_{AFT-LDA} = 0.52\), \(mean_{PCA-LDA} = 0.48\)), 1NN (\(p_{1NN} = 2.636\) – 14, \(mean_{AFT-1NN} = 0.46\), \(mean_{PCA-1NN} = 0.36\)) and DT (\(p_{DT} = 4.989e - 17\), \(mean_{AFT-DT} = 0.38\), \(mean_{PCA-DT} = 0.28\)) classifiers.

AFT also provides better results with fewer dimensions than PCA (54.98% accuracy with 45 dimensions for the former, versus 52.69% accuracy with 55 dimensions for the latter). The best results are achieved with a cluster size of 75 for AFT (maximum accuracy = 56.61%, \(e=0.48\), cluster size =75). The self-organising map results with a cluster size of 75 are depicted in Figure 1. This suggests that AFT features are able to capture the variations in the speech signal more accurately than MFCC and their PCA transformation.

From the confusion matrix of the best results of these experiments (Figure 5) it is observed that MFCC features provide less accurate results for all types of attitudes than PCA and AFT. The AFT provides better recall results than PCA but for neutral (PCA detected 51 instances (recall=51%) and AFT detected 50 instances correctly (recall=50%)) and friendliness (PCA and AFT detected 61 instances correctly (recall=60.40%)) there is almost no improvement in recall results. However the Amusement and Impatient recall results improves with AFT (37 instances of both are correctly detected (57% and 55.88% respectively)) features as compared to PCA (47 instances of both are correctly detected (47% and 46.08% respectively)). The AFT provides better precision results than PCA except only for Impatience where the AFT provides a precision of 48.45% for PCA and 47.50% for AFT. However the AFT provides the better F1,Score for all the attitudes than MFCC and ‘PCA transformation of MFCC’. Moreover, the MFCC features provide an overall accuracy of 43.07% with Kappa factor [20] of 0.32 and ‘PCA transformation of MFCC’ provides an overall accuracy of 52.69% with Kappa factor of 0.43. However, the AFT of MFCC provides the best results with an overall accuracy off 56.361% with Kappa factor of 0.48 as depicted in Figure 5.

We use a Venn diagram to visualise commonalities in classification of the best classifier (LDA) for each experiment. In Figure 6, the blue circle (labelled Target) represents the annotated labels, the yellow circle (Exp.2) represents labels predicted by LDA using the MFCC features, the green circle (Exp.1) represents labels predicted using the AFT of MFCC, and finally the red circle (Exp.3) represents labels predicted using the PCA transformation of MFCC. From the overlays shown in the Venn diagram, it is observed that there are 140 instances (18 of A, 27 of E, 21 of Fd, 22 of Fr, 29 of I and 23 of N) which have not been recognised by any of the feature set. However there are 160 instances (14 of A, 33 of E, 41 of Fd, 31 of Fr, 23 of I and 18 of N) which have been detected by all the three experiments. We also compared the predictive accuracies of our three best results using the mid-p-value McNemar test with a null hypotheses that, predicted labels of Exp.1, Exp.2 and Exp.3 have equal accuracy for predicting the target. The statistical test rejects the null hypotheses when compared with the results of MFCC features and transformed features (PCA and AFT) (\(p_{Exp.1-Exp.2} = 1.0284e - 05\) and \(p_{Exp.2-Exp.1} = 2.0194e - 08\) but is unable to reject the null hypothesis \(p_{Exp.2-Exp.3} = 0.1078\)) when the results of AFT and PCA are compared. This shows that although AFT provides better results than PCA on average, the difference is not statistically significant at the \(p < 0.05\) level for the best results. However, AFT provides the best F1,Score for all the attitudes, as depicted in Figure 5. This demonstrates the strength of AFT over MFCC and PCA transformation of features.

7. Acknowledgement

This research is supported by EU H2020 project SAAM under grant number 769661 at the University of Edinburgh, UK, and “ADAPT 13/RC/2106” project (http://www.adaptcentre.ie/) in the Design and Innovation Lab (DLab) at the School of Computer Science and Statistics, Trinity College Dublin, the University of Dublin, Ireland.

6. Conclusion

A novel Active Feature Transformation (AFT) method has been proposed and used in an attitude recognition task. The results show that the AFT method is able to effectively reduce the dimensionality of MFCC features for attitude recognition. It outperformed the statistical-functional representation (MFCC Features) of features for speech segments and also outperformed principal component analysis in terms of accuracy and dimensionality reduction using linear discriminant analysis, 1-nearest neighbour and decision tree classifiers. In future we intend to evaluate the performance of the AFT method for multiple feature sets, including prosodic, voice quality, EEG, and image features on multiple prediction problems such as sound events detection, speaker recognition, emotion recognition, human action recognition and dementia recognition.
8. References


