Aggression Detection in Speech Using Sensor and Semantic Information

No Author Given

No Institute Given

Abstract. By analyzing a multimodal (audio-visual) database with aggressive incidents in trains, we have observed that there are no trivial fusion algorithms to successfully predict multimodal aggression based on unimodal sensor inputs. We proposed a fusion framework that contains a set of intermediate level variables (meta-features) between the low level sensor features and the multimodal aggression detection [1]. In this paper we predict the multimodal level of aggression and two of the meta-features: context and semantics. We do this based on the audio stream, from which we extract both nonverbal and verbal information. Given the spontaneus nature of speech in the database, we rely on a keyword spotting approach in the case of verbal information. We have found the existence of 6 semantic groups of keywords that have a positive influence on the prediction of aggression and of the two meta-features.

Key words: emotional words spotting, aggression detection, multimodal fusion

1 Introduction

Our project addresses the problem of automatic audio-visual fusion in the context of aggression. In particular, we are analyzing a database with audio-visual recordings of aggressive and unwanted behavior in trains. The level of aggression in audio, video and both simultaneously has been annotated on a 3 point scale. Having these three types of annotation we explore methods of predicting the multimodal label given the audio and video labels. We find that there is a large diversity of combinations of the audio, video and combined sensor data, and no trivial fusion method based on simple rules or a classifier has a good performance in predicting the multimodal label given the unimodal ones. We observe that there are a number of concepts inherent in the multimodal data that are missed when only unimodal information is assessed. We propose to use an intermediate level in the fusion framework. This level contains the audio and video predictions and a set of five high level concepts (meta-features) that have an impact on the fusion process: Audio-Focus, Video-Focus, History, Context and Semantics.

In the context of this fusion framework, so far our predictions were based on low level sensor features from audio (nonverbal) and video [1]. But the audio modality has two components: nonverbal (prosody - how things are said), and

verbal (what is being said). In this paper we focus of both of these components to predict aggression. As opposed to our previous work, we add semantic information based on the linguistic content by extracting aggression related keywords. From the spoken text we discovered six classes of words which can be useful in the prediction of aggression. We use these linguistic features to enhance multimodal aggression detection. Furthermore, we show that the features are good predictors for the Context and Semantics meta-features which are introduced in section 4 and which are important in our final multimodal assessment.

This paper is organized as follows. In section 2 we provide an overview of related work. We continue with a description of the database and how it was annotated. In section 4 we present our fusion model and prove that it is beneficial. Details on the acoustic and linguistic features are given in section 5. In section 6 we describe the approach and results for multimodal aggression prediction based on the audio verbal and nonverbal features, and in section 7 we use the same features to predict the Context and Semantics meta-features. The paper ends with conclusions and directions for future work.

2 Related work

Research in the field of audio-visual fusion is getting more and more attention [2]. However, in the context of surveillance applications, the focus is mostly on video only, or on the acoustic component of sound, while linguistic information is mostly ignored. As we will show, the linguistic component provides a lot of relevant information.

In the field of emotion recognition, the linguistic component was successfully combined with prosodic information in [5]. Their approach was to use a Belief Network for spotting of emotional key-phrases, based on their frequency of appearance in emotional utterances. A string-based audio-visual fusion method for dimensional affect assessment is proposed in [3]. Besides head gestures, Action Units (AUs) and non-verbal acoustic events, they have used a keyword detection algorithm. The words relevant for each affect dimensions are detected by correlation based feature subset selection. The Dictionary of Affect in Language, [6] contains valence and arousal scores for a large dataset of words. However, we decided not to use it because we observed that our database contains many bad words and expressions related to aggression that are not in the database.

A different application of audio-visual fusion is event detection in team sports videos [7]. The use of semantic information from external sources like match reports or real-time game logs in detecting events proved to be a key contribution in their approach.

3 Dataset and human assessment

3.1 Database of aggression in trains and its annotation

We defined a set of rules that describe normal behavior in trains. A set of 21 scenarios were generated, each of them breaking one or more of the rules of normal

behavior. The scenarios contain different abnormal behaviors like harassment, hooligans, theft, begging, football supporters, medical emergency, traveling without ticket, irritation, passing through a crowd of people, rude behavior towards a mother with baby, invading personal space, entering the train with a ladder while the conductor is against, mobile phone harassment, lost wallet, fight for using the public phone, mocking a disoriented foreign traveler and irritated people waiting at the counter or toilet. The scenarios were performed freely by a team of semi-professional actors to ensure realistic outcome. The total length of the audio-visual recordings is 43 minutes.

The level of aggression has been annotated on a three point scale (1-low, 2-medium, 3-high). The annotation has been done in three settings: audio only, video only and looking at and hearing the data simultaneously. For each annotation scheme the data has been split into segments of homogeneous aggression level by two expert annotators. The data is unbalanced, dominated by neutral samples.

3.2 Analysis of human assessment

We want to understand how the audio, video and multimodal annotations relate to each other. Especially we are interested in those cases where these three labels do not agree. We have computed a 3D confusion matrix of the annotations, which we call a confusion cube. The axes of the cube correspond to the three types of annotation, audio-only (A), video-only (V) and multimodal (MM).

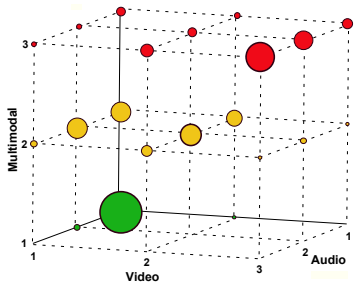

Fig. 1. Confusion cube of audio, video and multimodal annotations.

Given the difficulty of coming up with a multimodal label by having the audio and video labels, we divide the samples from confusion cube in two groups:

- On-diagonal - audio, video and multimodal are equal.
- Off-diagonal - for these points the multimodal label is not equal to at least one of the two unimodal labels. These are the samples that we consider challenging and on which we will base the following section. Note that 46% of the data falls in the off-diagonal case.

4 Fusion model based on meta-information

In order to understand why there is no straightforward relation between the multimodal label and the unimodal labels we have carefully inspected the off-diagonal samples. It occurred that there are a number of intermediate factors that are not obvious from one modality only. We call these factors meta-features and we have identified a set of five that have an influence on fusion as follows:

- **Audio-focus (AF)** - the rater was more influenced by the audio channel than by the video channel in his/her final assessment of the multimodal level of aggression.
- **Video-focus (VF)** - the video modality had the final impact on the multimodal assessment. The two meta-features are mutually exclusive. It can also be the case that none of them is active, when there is no dominant focus on one modality.
- **Context (C)** - the of aggression is strongly influenced by the region of interest or situations in which the actions is taking place and what is appropriate or not in that case.
- **History (H)** - illustrates the effect that negative events can have over time. Such a feature is not captured in the unimodal annotation, since we asked annotators to rate each segment on purely what it suggested. However, this was included in the multimodal annotation.
- **Semantics (S)** - the semantic interpretation of the scene is pointing to abnormal behavior.

We have annotated the meta-features on the samples of the cube which were off-diagonal. The most frequently activated meta-feature was Semantics. It was annotated in 61% of the off-diagonal samples. The other four meta-features were present in percents of 24% Audio-Focus, 22% Video-Focus, 25% History, and 16% Context.

The proposed fusion model is depicted in Figure 2. Instead of trying to predict the multimodal label from the low level features, we propose to use an intermediate level. The picture shows the contributions of audio, video (depicted in green) as well as the meta-features (in orange) in constructing a multimodal output. It also shows that the previous audio and video assessments influence the History variable. The AF, VF and H meta-features are predicted based on low sensor features. For the prediction of Context and Semantics, we needed to add semantic information. The focus of the paper is to use the audio modality in terms of prosody and words to predict the level of aggression and the Context and Semantics meta-features. This is illustrated in the figure by solid lines.

To verify the effect of the meta-features on multimodal fusion we have trained a Random Forest (RF) classifier on the Audio and Video labels and predicted the Multimodal label. We then added the meta-features to the feature vector, and the performance improved from 86% to 92% weighted average. Class 3, which corresponds to the highest level of aggression and which is less represented in the data and harder to predict has a significant improvement of 18% absolute.

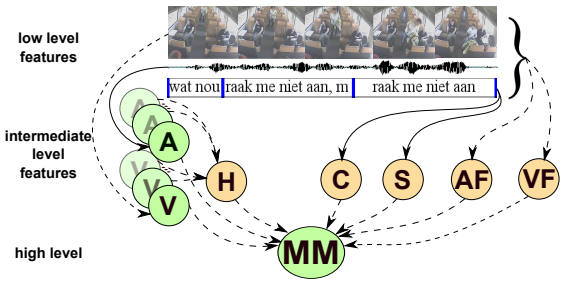

Fig. 2. Fusion model based on meta-features. The dashed lines represent human annotations of the data streams. The solid lines are learned by classifiers.

More information can be found in [1]. In the enxt sections we continue with a semantic analysis of speech to compute Context, Semantics and their influence on the Multimodal label.

5 Acoustic and linguistic features

5.1 Acoustic features

Vocal manifestations of aggression are dominated by negative emotions such as anger and fear, or stress. The audio feature set consists of 30 features inspired from [4]: speech duration, statistics (mean, standard deviation, slope, range) on pitch (F0) and intensity, mean formants F1-F4 and their bandwidth, jitter, shimmer, high frequency energy, HNR, Hammarberg index, center of gravity and skewness of the spectrum. These features are computed on segments of length equal to 2 seconds, because this resembles better what we can expect in real-time processing.

5.2 Linguistic features

The used language has the characteristics of spontaneous speech, with a lot of interruptions, restarts, overlapping speech, slang, interjections and nonverbal utterances('um','eh','he'). The first step was to manually transcribe the database to text. Given the nature of the data no natural language processing approaches can be used successfully, since most utterances are not grammatically correct. A keyword spotting approach was used instead. To start with, we want to see what is the added values of linguistic analysis in perfect conditions, without being affected by transcription errors.

We define aggressive keywords as words or expressions that convey aggressive states or that are stimulating aggressive states. These keywords were selected manually from the transcription by three annotators. They have been clustered in 6 classes based on expert knowledge:

1. **positive emotions**: this class contains words that express positive emotions, e.g. 'nice', 'cool', 'helpful', 'everything's under control'.
2. **negative emotions**: this class contains words/expressions conveying negative emotions, e.g. 'irritated', 'don't want to', 'unfair', 'disturb', 'lousy'.
3. **actions**: this class contains words/ expression related to actions that relate to the fact that somebody is being disturbing, e.g. 'don't touch me', 'behave normally', 'leave me alone', 'stay still', 'go away', 'pay attention', 'stop'.
4. **context**: the words in this class are good indicators of special contexts. e.g. 'police', 'ambulance', 'thief', 'drugs', 'sniffing', 'dead', 'criminal', 'wallet', 'arrested'.
5. **cursing**: this class contains cursing and offensive words.
6. **nonverbal**: we have added semantic tags to a number of nonverbal sounds, e.g. singing, clapping, knocking, noise and repetitions. These can also be detected automatically but in this paper we first wanted to determine their added value.

In Figure 3 we show the number of occurrences per word class in the database. Given the five classes, we have created a 6 dimensional feature vector with binary values given the presence or absence of words from the class.

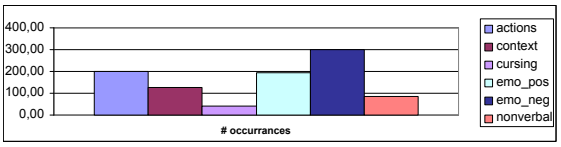

Fig. 3. Number of occurrences for each keyword class

6 Automatic prediction of multimodal aggression

In this experiment we are predicting the Multimodal label of aggression based on the prosodic and linguistic features. As a first step, we use the prosodic features to predict the Audio label. For this we use a logistic regression classifier. The posteriors of this classifier concatenated with the linguistic features form the final feature vector for Multimodal aggression prediction.

We have observed that the presence of our keywords does not have a strict local impact. Instead, the impact lasts for a longer time (in this case we consider a number of 2 seconds segments corresponding to the ones for which the prosodic features were computed). We test several configurations to find an optimal impact length.

With this setting, we compare two approaches. First, we use only the Audio posteriors to predict the Multimodal label. We then add the linguistic features to it, to illustrate their impact. In both cases we use a Random Forrest (RF) classifier.

A comparison of the performances of the RF classifier that predicts multimodal aggression using different impact length of the semantic words features is depicted in Figure 4. Because in a surveillance application we are interested in having the smallest miss rate for the aggressive cases, we choose the impact length of 11 segments as the most appropriate.

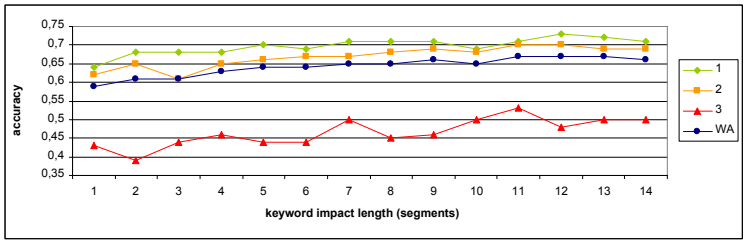

Fig. 4. Accuracies for prediction of the Multimodal label given different impact length of the keywords.

The results of using semantic keywords with an impact length of 11 segments shows an improvement of 10% absolute compared to using the Audio posteriors only. Furthermore, the improvement of class 3 (the most aggressive cases) is of 16% absolute. We did an information gain feature ranking and the most relevant linguistic features were context, negative emotions and nonverbal. Even though the linguistic features provide an improvement in predicting the Multimodal label, the performance is not high. However, the video modality was not taken into consideration and the Multimodal labels is based on that as well.

7 Automatic prediction of Context and Semantics

This section describes the prediction from low level features of the Context and Semantics meta-features from the intermediate level of our fusion framework depicted in Figure 2. The successful prediction of the meta-features has a positive influence of the final multimodal aggression assessment. For predicting the Context and Semantics meta-features we use the Audio posteriors concatenated with the linguistic features with the 11 segments impact length. In both cases we use a Support Vector Machine classifier with a second order polynomial kernel. Note that in this experiment we only use the subset of the data for which the three modalities do not agree (the off-diagonal of the confusion cube), because these are the most interesting cases for multimodal fusion.

The prediction accuracies for Context and Semantics are presented in Table 1. For both cases we have used a feature ranker based on information gain. In the case of Semantics, the most useful features were the posteriors of the Audio aggressive classes (2 and 3), actions, nonverbal and cursing. In the case of Context, the Audio posteriors had no influence. The top ranked features were nonverbal, context and cursing.

Class	Accuracy	
	Context	Semantics
0	0.95	0.50
1	0.69	0.87
WA	0.91	0.73

Table 1. Prediction accuracies for Context and Semantics for class 0 (meta-feature is not activated), class 1 (activated) and their weighted average (WA)

8 Conclusion and future work

In this paper we have used prosodic and linguistic information in the form of semantically meaningful keywords to predict aggression. We have done the experiment in the framework of our project focused on multimodal fusion for aggression detection. Besides predicting the level of aggression, we have shown that the linguistic information is a rich source for predicting Context and Semantics, two of the meta-features of our fusion framework. In our future work we will combine the linguistic, prosodic and video modalities to predict all variables in the intermediate level of our fusion framework. Based on them the multimodal level of aggression will be automatically assessed.

References

1. Anon. Reference ommited for blind review.
2. P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Springer Multimedia Systems Journal*, pages 345–379, 2010.
3. F. Eyben, M. Wollmer, M.F. Valstar, H. Gunes, B. Schuller, and M. Pantic. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 322 –329, march 2011.
4. I. Lefter, L.J.M. Rothkrantz, P. Wiggers, and D.A. Van Leeuwen. Emotion recognition from speech by combining databases and fusion of classifiers. In *Proceedings of the 13th international conference on Text, speech and dialogue*, TSD'10, pages 353–360, Berlin, Heidelberg, 2010. Springer-Verlag.
5. B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 1, pages I – 577–80 vol.1, may 2004.
6. C. M. Whissell. *The dictionary of affect in language*, volume 4, pages 113–131. Academic Press, 1989.
7. H. Xu and T.-S. Chua. The fusion of audio-visual features and external knowledge for event detection in team sports video. In *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, MIR '04, pages 127–134, New York, NY, USA, 2004. ACM.