Learning the Fusion of Audio and Video Aggression Assessment by Meta-Information from Human Annotations

Iulia Lefter [1,2,3], Gertjan J. Burghouts [2], Leon J.M. Rothkrantz [1,3]
[1] Delft University of Technology, Delft, The Netherlands
[2] TNO, The Hague, The Netherlands
[3] The Netherlands Defence Academy, Den Helder, The Netherlands
I.Lefter@tudelft.nl

***Abstract*—The focus of this paper is finding a method to predict aggression using a multimodal system, given multiple unimodal features. The mechanism underlying multimodal sensor fusion is complex and not completely clear. We try to understand the process of fusion and make it more transparent. As a case study we use a database with audio-visual recordings of aggressive behavior in trains. We have collected multi- and unimodal assessments by humans, who have given aggression scores on a 3 point scale. There are no trivial fusion steps to predict the multimodal labels from the unimodal labels. We propose an intermediate step to discover the structure in the fusion process. We call these meta-features and we find a set of five which have an impact on the fusion process. Using a propositional rule based learner we show the high positive impact of the meta-features on predicting the multimodal label for the complex situations in which the labels for audio, video and multimodal do not reinforce each other. We continue with an experiment by which we prove the added value of such an approach on the whole data set.**

Keywords-multimodal fusion, aggression, meta-features.

I. INTRODUCTION

Historically multimodal information fusion was seen as a solution to improve the overall performance of a system [1]. Many sensor fusion models have been developed. In the current paper we develop a new model inspired by the information fusion by humans. In the case of fusion of audio and video data, we expect an accurate multimodal prediction with high confidence outcome in the case when the two modalities indicate the same result. But the strength of fusion lies in improving the prediction in cases in which one of the modalities is lacking, or when one modality has a lower accuracy, or when the modalities do not reinforce each other. These are the cases which make the fusion process more complex and the cases on which we will focus in this paper.

We are aiming at improving multimodal aggression detection. The concept of aggression involves more complex behaviors than just pure violence like physical fights. We can distinguish between affective aggression, instrumental aggression, behavior that is considered unwanted because it is not appropriate given a specific context. For example people are expected to have a specific behavior in a train – entering, sitting, reading, showing their ticket to the conductor, leaving. When instead they are running, moving a lot, making a lot of noise, not showing the ticket to the conductor, disturbing the other passengers, these are considered deviant behaviors. In most of these cases, the automatic detection of aggression becomes harder because there is no pure violence, and the fusion of modalities plays a central role. In the case of aggression in trains, we find that humans exploit all information they can infer from both modalities and that they use context-enhanced semantic interpretation of what they see and hear.

We consider a large audio-visual database of train incidents which include aggression in various degrees and normal, neutral situations. The aggression level was annotated on a 3 point scale in 3 different setting: the raters are presented with audio data only (they don't see the video stream), with video data only (they don't hear the audio stream) and finally with multimodal data (audio and video simultaneously). Having these three types of annotation we explore methods of predicting the multimodal label given the audio and video labels. Our expectation is that similar procedures can be used in the case of fusing unimodal sensor data to predict multimodal data. We find that there is a large diversity of combinations of the three, and no trivial fusion method like simple rules or a classifier has a good performance in predicting the multimodal label given the unimodal ones.

In this paper we present our analysis of the samples for which the audio, video and multimodal labels do not agree while being provided with the multimodal data. We observe that there are a number of concepts inherent in the multimodal data that are missed when only unimodal information is assessed. Instead of trying to predict multimodal based on the two unimodal assessments, we propose to use an intermediate level. This level contains a set of five high level concepts that have an impact on the fusion process. Throughout the paper we refer to these concepts as meta-features. These concepts are *Audio-Focus*, *Video-Focus*, *History*, *Context* and *Semantics*. In themselves, they do not convey information about the multimodal level of aggression, but they give more insight into how the information from audio-only and video-only streams can be fused better. Throughout the paper we will show the

positive impact of these concepts in two experiments for which the meta-features have been manually annotated.

Our main contributions are:

- Identifying five high level concepts ('meta-features') that have a significant impact in the multimodal fusion process as done by humans.
- Proposing a computational fusion algorithm based on rules which benefits of the five meta-features.
- Testing the approach on realistic data and predicting unseen situations.
- Analyzing more subtle forms of aggression than just pure violence.
- Improving the distinction between intermediate and severe aggression by the proposed fusion algorithm.

This paper is organized as follows. Section II contains a review of related work. In Section III we give a detailed view of the dataset we have used, how it was annotated. In Section IV we prove the complexity of the fusion problem by examining a 3D confusion matrix of the 3 labels. We continue in Section V with the description of the proposed meta-features and the fusion model. We show in Sections VI and VII that the meta-features have added value for the prediction of multimodal aggression by two experiments. The first experiment (Section VI) is based on the data samples for which the three modalities do not reinforce each other, and the second experiment (Section VII) proves the beneficial impact of the meta-features on the whole dataset. The paper ends with our conclusion and directions for future work.

II. Related work

There is a considerable amount of work focusing on multimodal fusion as illustrated by Atrey et al. [1]. We highlight a few of them that are related to aggression detection. Furthermore, there are a number of contributions which use intermediate levels in the fusion approach, which we will also address, as this is the focus of our paper.

A two stage approach for detecting fight based on acoustic and optical sensor data in urban environments is presented by Andersson et al. [2]. In the first stage low level features are used to recognize a set of intermediate events. The events based on video data are related to the behavior of crowds: normal activity, intensive activity by a few or by many persons, small crowd or large crowd. From sound it is distinguished between low sounds and high sounds. In [3] the focus is on detecting violent scenes in videos for protection of sensitive social groups by audio-visual data analysis. The audio stream is processed and a set of special classes are detected, like music, speech, shots, fights and screams. From video the amount of activity in the scene was used to discriminate between inactivity, activities without erratic motion and activities with erratic motion (fighting, falling). An audio-video aggression detector for train stations is described by Zajdel et al. [4]. At the low level they perform analysis of the audio and video stream to detect events like scream, passing train or articulation energy. The fusion is done using a Dynamic Bayesian Network. These three papers, [2-4], share some similarity with our approach since they focus on surveillance related topics and use an intermediate step to get from the low level features to multimodal assessment. While their focus is on pure violence, we focus on more subtle behaviors as well. Detecting such behaviors is important since they are aggressive or disturbing in nature, even if there is no explicit aggression. Also, these are the cases that can lead to more serious aggression so they are likely to help prevention if detected in an early stage. The papers [2-4] use an extra level to detect aggression related events. Our meta-features are different from that: they do not confer information about aggression but guide the fusion process. Another difference is that the events from [2-4] are detected from the unimodal streams, while our meta-features capture higher level information from the multimodal input.

A number of papers illustrated the use of extra information that is not in the unimodal streams in the process of fusion. A probabilistic fusion framework for photo annotation using metadata of multiple modalities is presented in Wu et al. [5]. They use contextual information (location, time, and camera parameters), visual content and a semantic ontology. The causalities between variables and semantic labels are modeled using influence diagrams. Detecting events in sports has been studied in [6]. The idea is to use external knowledge sources such as match reports and real-time logs together with the audio-visual features. Our research resembles [5-6] since we are also interested in context information, yet specific for surveillance and security. Quality based fusion for biometrics is presented in [7]. Here the quality of each sensor is regarded as a meta-feature (how trustworthy each sensor is) and it is used as a feature in the fusion process. The similarity between their approach and ours is that the power of the meta-features lies in influencing the fusion process. The domain of application and the type of meta-information are however different: detection of unwanted situations in surveillance will require other types and sources of context. We explain this in the following sections.

III. Dataset and Human Assessments

We want to understand aggression and its appearances in videos. Therefore, we have chosen a bottom-up or data-driven approach. This is the reason why the video database with incidents and normal situations, which we use throughout this paper, has a central role (Section IIIA). Further, the human uni- and multimodal annotations for this dataset are essential to understand aggression and how humans fuse information (Section IIIB).

A. Database of aggression in trains

Aggression and unwanted behavior have a variety of manifestations. To distinguish those, we considered a baseline of normal behavior. We defined a set of rules that describe normal behavior in trains. Passengers are expected to give priority to other passengers leaving the train before entering. Next, they are supposed to go quietly to an empty seat, store their luggage, do not disturb and do not intrude personal space of other travelers. While using the mobile phone one should speak quietly. At the public phone, public toilet and at the

counter they should be patient and polite. They should keep to their personal belongings. In the presence of the conductor they should follow his directions. Violence of any kind, smoking and consumption of alcohol are prohibited. Furthermore, the travelers are expected to be polite and mind people that might need a seat more.

A set of 21 scenarios were generated, each of them breaking one or more of the rules above. 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. Details about the database are described by Yang [8].

These scenarios were performed by a team of actors. The actors were given scenario descriptions in terms of storyboards. In this way they had a lot of freedom to play and interpret the scenarios which assures realistic outcomes. The total length of the recordings is 43 minutes. Examples are shown in Figure 1. The lighting conditions, the number of people, the variety of the situations are all very different. That makes this a challenging dataset, appropriate for analyses of aggression, and a good benchmark for our fusion method.

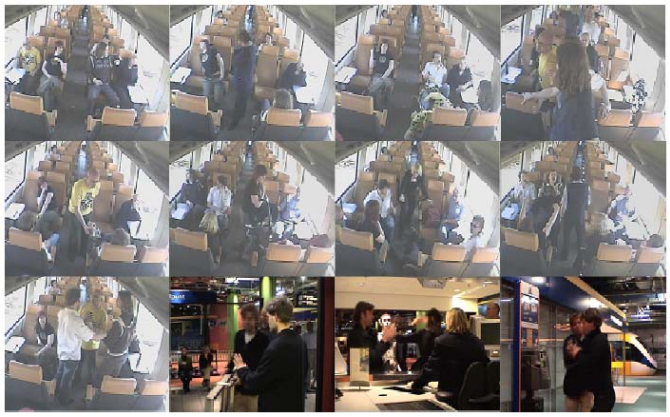

Figure 1. Examples of scenarios from the train database. The behavior variations may be subtle whereas the recording conditions vary significantly.

B. *Human annotation*

The human annotation has been done in the following settings:

- audio-only - the rater is listening to samples of the database without seeing the video,
- video-only - the rater is watching samples of the database without sound, and
- multimodal - the rater used both video and audio samples.

1) Segmentation

For each annotation scheme the data has been split in segments of homogeneous aggression level by two expert annotators. In general, there was a finer segmentation for the *Audio* - a mean duration of 8.5 seconds, and a coarser one for video and multimodal with mean segment durations of 15 and 16 seconds respectively. The different segment durations are inherent in the data and in the way each modality is dominant or not for a time interval. In the case of audio the resulting segment durations are shorter. That is, when people are taking turns to speak, the aggression level changes with the speaker, so the aggression is more fine-grained. As a final step the segmentation was corrected by the first author such that small variations for starting and ending a segment in different modalities were removed. When annotating audio-only and video-only the raters were asked to consider strictly the segment they are watching or hearing. When rating the multimodal data they were allowed to use any cue they could get from the two modalities.

2) Annotation

For each segment we asked the raters to imagine that they are operators watching and / or listening to the data. They had to rate each segment on a 3 point scale as follows:

- label 1 - normal situation,
- label 2 - medium level of aggression / abnormality (the operator's attention is drawn by the data), and
- label 3 - high level of aggression / abnormality (the operator feels the need to react).

Seven annotators rated the data for each setting (modality). The inter-rater agreement is computed in terms of Krippendorff's alpha for ordinal data. The highest value is achieved for audio, namely 0.77, while video and multimodal are almost the same, 0.62 and 0.63 respectively. One reason for the lower interrater values for video and multimodal, compared with audio, can be the finer segmentation that was achieved for audio, but also that raters perceived verbal aggression in very similar ways due to the semantics in the wording and explicit aggression in sounds. The interrater values do not reflect perfect agreement but are reasonable given the task.

3) Labels

Figure 2 displays distribution of the labels in terms of duration for each annotation setting. It can be noticed that the data is unbalanced, as there are mostly neutral samples (label 1). Further, the duration of the segments with labels 2 and 3 is longer in the case of multimodal annotation. This can be caused by the additional context information that people get when using an extra modality and from the more accurate semantic interpretation that they are able to give to a scene, even if it does not look or sound extremely aggressive.

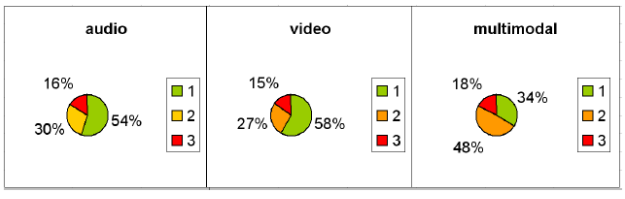

Figure 2. The duration of each class based on the three different types of annotations. Class 3 incidents occur the least while they are the most important to detect. Note: the class 3 labels are often not overlapping in the three modalities (see text).

The relative scarcity of very aggressive cases (label 3) makes the problem addressed in this paper a challenging one. Important: even though it seems from the pictures that there are about the same amounts of samples with label 3 in all modalities, they are not always the same cases, as we will show in the next section.

IV. Analyses of Human Assessments

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. Typically, for this, one would consider a confusion matrix for analyzing pair-wise (tuple) confusions. We have three modalities and want to go one step deeper into analysis of triplets. In Figure 3 we have plotted 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). The sizes of the colored circles are proportional to the numbers in the confusion cube. The confusion cube represents the triplets of labels by dots. The green dots correspond to the cases when MM=1. We can see that when MM=1, in almost every case A=1 and V=1. The yellow dots and red dots represent the cases when the aggression level assessed based on multimodal information is 2 and 3 respectively. For these two particular cases we can observe that there are many combinations of A and V possible that have multimodal labels 2 and 3. A trivial fusion scheme is therefore not possible.

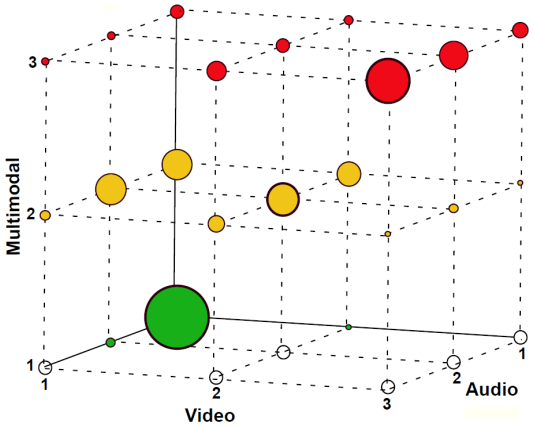

Figure 3. Confusion cube of audio, video and multimodal annotations. Note that 46% of the cases are not on the diagonal, which means that often the labels in the three modalities do not agree and that makes the prediction of the multimodal labels hard.

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:

1. *On-diagonal* - the points on the diagonal of the confusion cube - the points (1,1,1), (2,2,2) and (3,3,3) corresponds to the cases when the labels of audio, video and multimodal are equal. These are the cases in which the modalities reinforce each other, the easy samples (with stronger contour in Figure 3).
2. *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.

It is interesting to note that 46% of the data falls in the off-diagonal case. This gives us even more incentive to investigate the fusion process of those cases.

V. Decomposition of Aggression into Meta-Features

A. Model

In order to understand why there is not a straightforward relation between the multimodal label and the unimodal labels we have carefully inspected those 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:

1) *Audio-focus (AF)*

Audio-Focus means that 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. It is important that this feature is not confused with focus of attention in general. It can be the case that one of the modalities is more dominant, for example when two people are having a heavy discussion, audio is more dominant than video. However, this is not what this feature is about. *Audio-Focus,* in our case, can be different from focus of attention. For example in a scene with a lot of movement and pushing, about which from video only it can be concluded that there is a high level of aggression, but from audio you can realize that this is about a bunch of friends. When the multimodal assessment for such a case is a low level of aggression because of the final influence of audio, we say that the *Audio-Focus* feature is active.

2) *Video-focus (VF)*

This feature is the opposite of *Audio-Focus*. It means that 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 not of them is active, when there is no dominant focus on one modality.

3) *Context (C)*

Many times we are dealing with assessing situations that are not in themselves aggressive. For instance running is not an aggressive act. But when we confer a specific context to that action, and put it in the corridor of the train, the action becomes suspicious. The *Context* meta-feature is a tag that accompanies situations whose level of aggression is strongly influenced by the region of interest it is taking place in, about what is appropriate or not in that case. For example in the train compartment there are clear expectations regarding what passengers do in the corridor, on the seats, at the entrance. The presence of people with a special role, in our case it can be the conductor, also influences the context by implying an expectation of specific behavior, e.g. show tickets.

4) History (H)

This is a meta-feature that illustrates the effect that negative events can have over time. Such a feature is hard to be captured in the unimodal annotation, since each segment was purely annotated based on what it suggested. Nevertheless, this is not a realistic assumption for the multimodal case. When one of the passengers is fainting and falls on the ground, this is obvious from all modalities. However, afterwards we cannot find evidence of abnormal behavior in video, but in the multimodal assessment, that fact that we know that a person fell and it was not helped triggers a high level of abnormal behavior.

5) Semantics (S)

The last meta-feature is concerned with the cases in which there is no pure violence, but the semantic interpretation of the scene is pointing to abnormal behavior. This is the case for instance in a segment when a man harassing is a woman who obviously does not want to have anything to do with him. It is hard also for humans to make a very sharp distinction between context and semantics. We do not claim to have achieved this perfectly.

B. Fusion scheme

The proposed fusion model is depicted in Figure 4. It shows the contributions of audio, video as well as the meta-features in constructing a multimodal output. It also shows that the previous multimodal assessments influence the *History* variable. The colors in the figure make two distinctions. First, the variables in green represent the initial configuration of such a fusion system (before the meta-features are added). Second, the entities in green represent variables that indicate the level of aggression given some input. On the other hand, the meta-features, colored in orange, are not related to the level of aggression. They have only an influence on the fusion process and have by themselves no discriminative power.

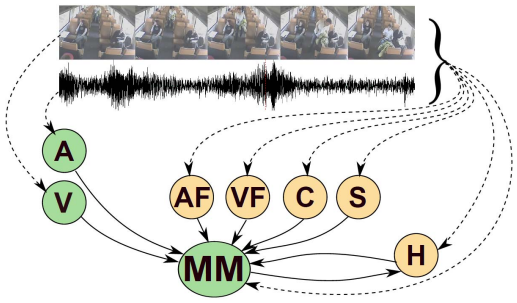

Figure 4. Fusion model based on meta-features. The dashed lines represent human annotations of the data streams. The solid lines are learned by the fusion method.

C. Annotation of the meta-features

We have annotated the meta-features on the samples of the cube which were off-diagonal. This means that we looked again at the data, and for each segment where MM was different that A and/or V, we have annotated the activated meta-feature(s). In the majority of the case one or two meta-features were annotated.

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*.

In Table I we present the occurrences (in %) of each meta-feature given the different combinations of A, V and MM. The percentages higher than 10 are printed in bold. We observe that for a number of A, V and MM configurations we obtain the largest percentages of meta-features. We see that *History* and *Semantics* have a strong influence in cases where Audio and Video are both 1, but Multimodal is 2. So their impact is increasing the multimodal assessment as compared to A and V.

Audio-Focus is mostly encountered in the cases where A and V are 2 and 1 respectively (70% of the cases in which AF was annotated). It has the effect of making the Audio label take over the Video assessment. In this case the "winning" label for MM is 2, corresponding to the highest aggression prediction between A and V. The same effect is noticed in only 10% of the cases, when Audio is 2, Video is 3 and MM is 2. Audio is again dominant but this time it corresponds to less aggression than Video. Similarly, *Video-Focus* occurs mostly in cases where A=1, V=2 and MM=2 (48,8% of the cases in which VF was annotated).

We observe that *History* appears mostly when MM is higher or equal to the maximum between A and V. The *Context* meta-feature occurs the most in the A=2, V=1, MM=2 case. We can translate this to "it doesn't look aggressive (V=1), but given the context it is not appropriate". Semantic is dominant in the cases where A=1, V=1 and MM=2, so it is used to amplify the aggression level, but also to lower it in the case of A=3, V=2 and MM=2.

TABLE I. OCCURRENCES OF THE META-FEATURES GIVEN DIFFERENT CONFIGURATION OF AUDIO,VIDEO AND MULTIMODAL PREDICTION IN %

A	V	MM	AF	VF	C	H	S
1	1	2			7,3	**31,1**	**26,5**
1	1	3				7,4	2,8
1	2	1			2,1		0,6
1	2	2		**48,8**	**21,9**	**15,5**	**19,5**
1	2	3				3,4	1,1
1	3	2		2,4	6,3	0,7	0,8
1	3	3		0,8		8,1	3,3
2	1	1		5,5			1,9
2	1	2	**70,0**	3,1	**51,0**	3,4	**10,3**
2	1	3	0,7			1,4	0,6
2	2	3	1,4			6,8	1,9
2	3	2	**10,0**				
2	3	3	0,7	**11,8**		**14,9**	6,4
3	1	2	2,1	1,6	4,2		4,2
3	1	3	2,9		2,1		1,1
3	2	2		**26,0**	2,1		**13,1**

VI. EXPERIMENT 1 – THE EFFECT OF USING META-FEATURES FOR MULTIMODAL PREDICTION IN THE OFF-DIAGONAL CASE

A. Experiment setup

This experiment is based only on the data that is off-diagonal. The input consists of the human annotated labels for *Video*, *Audio* (both on a 3 point scale) and the five meta-features (with binary values). The goal of the experiment is to explore if we can predict the *MultiModal* label (3 point scale)

with a higher accuracy when using the meta-features as opposed to when trying to predict the multimodal label only from audio and video.

For processing reasons, the initial segmentations based on segments with equal levels of aggression were transformed into finer segmentations. Each segment has a length of 2 seconds.

Model learning and prediction were done using a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by Cohen [10]. We have opted for a rule-based or tree-based learning as opposed to other statistical classifiers like support vector machines, because the outcomes are very easy to visualize, understand and interpret. In particular we have chosen the RIPPER algorithm due to its high accuracy (better than the C4.5 decision tree) and its capability of outputting a very condensed rule base model. The pseudo-code in Fig. 4 shows the seven if-then-else rules obtained by training RIPPER on our data.

```
if (VF >= 1) and (V <= 1) and (A <= 2) then MM=1
  else if (V >= 3) and (H >= 1) then MM=3
    else if (V >= 2) and (A >= 2) and (H >= 1) then MM=3
     else if (V >= 3) and (AF <= 0) and (C <= 0) then MM=3
       else if (A >= 3) and (AF >= 1) then MM=3
         else if (H >= 1) and (S >= 1) and (VF <= 0) then MM=3
           else MM=2
```

Figure 5. RIPPER model for MultiModal prediction from Audio, Video and meta-features labels.

The rules show the effect of the meta-features in the fusion process. Take for example the third rule. This rule means that whenever audio and video have value of 2 or 3, the actual aggression level in multimodal would be higher (3) because there is an effect of an event in the past.

The experiment was done with 10-fold cross-validation framework.

B. *Results*

We discuss the results in terms of confusion matrices since they provide us insight in confusions between the levels of aggression (1-3). Note that in the off-diagonal case class 1 is poorly represented (as can also be seen from the confusion cube). This is why it is harder to predict. Nevertheless, we pay special attention to classes 2 and 3 since these are more interesting for us. Class 2 contains the less aggressive cases but which are very important to detect since they might lead to something more serious. Class 3 is the most important not to miss, and due to the unbalanced data also the hardest to predict, see Lefter et al. [9].

With the meta-features, the results show an improvement in class 3 from 57% to 82%, so an improvement of 25%. Class 2 suffers a minor loss of 1%. The recognition of class 1 (only 13 instances) improves from 0 to 30%. The overall weighted average accuracy increases from 85% to 91%, while the unweighted average increases from 50% to 70%. Note that given the unbalanced data we are dealing with, the unweighted average is a more appropriate performance measure. And since in an aggressive behavior detector we value the most having a small miss rate for the most aggressive cases, we are interested mainly in the performance of class 3 (given .

TABLE II. CONFUSION MATRICES IN % FOR MULTIMODAL AGGRESSION PREDICTION USING AUDIO AND VIDEO (LEFT) AND AUDIO, VIDEO AND THE META-FEATURES (RIGHT).

	Predicted		
True	1	2	3
1	**0**	0	0
2	100	**95,6**	42,7
3	0	4,4	**57,3**

	Predicted		
True	1	2	3
1	**31**	0.9	0
2	69	**94.6**	14.3
3	0	4.5	**85.7**

It is interesting to see how much each meta-feature contributes to the final improvement. Iteratively, we have added the next most informative meta-feature. If only one meta-feature is used, then the most informative is *Video-Focus*, but the best combination of two meta-features was *Audio-Focus* and *History*. The results are illustrated in Figure 6 in terms of F1 measure per class and weighted average. The size of the marker is proportional to the amount of data that falls in that category.

Apparently the use of *Audio-Focus*, *History* and *Semantics* has the most benefits for class 3. *Context* is the last meta-feature added. It does not have a positive influence on class 3; however it is beneficial for the overall performance.

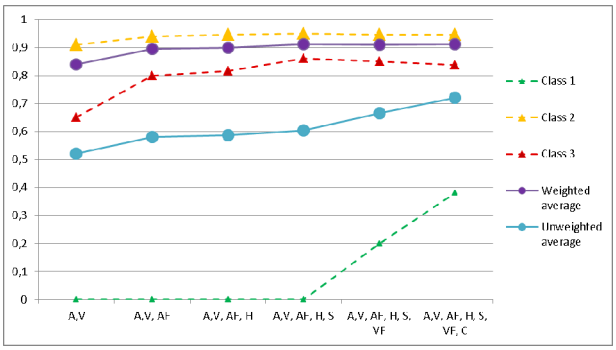

Figure 6. Added value of the meta-features in terms of F1 measure for each class (dashed lines) and their weighted and inweighted average (solid lines).

VII. EXPERIMENT 2 – THE EFFECT OF USING META-FEATURES FOR MULTIMODAL PREDICTION IN ALL CASES

A. *Experiment setup*

As opposed to the previous experiment, we now use the entire database (both off- and on-diagonal samples). The goal is to explore if we can predict the multimodal label (valued on a 3 point scale) with a higher accuracy when using the meta-features as opposed to when trying to predict the multimodal label only from audio and video also in the case when we use entire data. Our initial idea of splitting the data into the on- and off-diagonal subsets was to understand multimodal fusion in the 46% cases when the labels do not agree. Based on that, we had two models:

1. On-diagonal: the model is easy, MM=A=V so we take MM=A.

2. Off-diagonal: the model is more complex and the prediction of MM is based on rules about A, V and the meta-features (note that A=V may still hold).

At prediction level we are able to compute the values for A and V. Based on them we have to decide which model to choose. When A and V are different we always choose the off-diagonal mode. However when A and V are equal it is not clear which model to use. It can be that the correct model is the on-diagonal one, and then MM=A, but it can also be that given the influence of some meta-features MM is different than A and V. For these special cases we can not predict correctly which model to choose. We can assume that a system would opt for the easiest model and we loose some accuracy. In this experiment we are concerned with the performance of the system given this setting. Can we still gain accuracy on the entire dataset with the meta-features compared to the standard setting of using just A and V?

B. Results

In this section we compare results for three settings. The first confusion matrix (in Table III, left) summarizes the baseline results: here MM is predicted only from A and V. The second confusion matrix (right) describes the results for including the meta-features in the ideal case: for each sample we know which prediction model to use. We notice improvements especially for the aggressive classes 2 and 3 of 23 and 16% respectively.

TABLE III. CONFUSION MATRICES IN % FOR THE BASELINE CASE (LEFT) AND IDEAL CASE (RIGHT) .

	Predicted		
True	1	2	3
1	**97**	22.1	4.9
2	3	**73.8**	22.1
3	0	4,1	**73**

	Predicted		
True	1	2	3
1	**97.7**	1	0
2	2.3	**96.4**	10.2
3	0	2.6	**89.8**

The confusion matrix in Table IV shows the results for the worst case scenario: for the samples with A=V we don't know which model to use and therefore always use the easy on-diagonal model. The performance drops, but we can still achieve an improvement of 10% for the most important class, class 3. The other classes do not improve the baseline but also do not suffer any loss.

TABLE IV. CONFUSION MATRIX IN % FOR THE WORST CASE SCENARIO.

	Predicted		
True	1	2	3
1	**97.7**	23.1	4.9
2	2.3	**74.9**	11.9
3	0	1.95	**83.2**

VIII. CONCLUSION AND FUTURE WORK

We have addressed the problem of understanding and improving automatic fusion of audio and video cues in the context of aggression detection. In order to do that, we have considered a database with recordings of a variety of aggressive and unwanted behaviors in a train setting. The data is very well suited for our analysis since it is very rich in context and semantic content, as opposed to other datasets that only contain pure physical violence.

Our analysis showed that there are more aspects that have a key impact in the way how humans reason and come up with an assessment of aggression based on multimodal data. We have identified such aspects and we have proposed a set of five meta-features to capture them.

We have validated the meta-features by two experiments. We show that by using them we can achieve better prediction accuracy in the multimodal case, especially for the most important and hard to predict class, namely of the highest degree of aggression and subtle unwanted situations. The use of meta-features makes the process of fusion transparent and the set of rules employed by the rule based classifier show this computationally.

Future work will consider the estimation of audio, video and the meta-features computationally from low level audio and visual features. We plan to predict Audio, Video, *Audio-Focus* and *Video-Focus* based on state-of-the-art audio and visual features. For *History* we plan to asses the behavior of audio and video over a time-window. For *Semantics* we will use the linguistic content. *Context* can be computed starting from dividing the video image into regions of interest and recognizing events from a predefined list of normal and abnormal behaviors.

REFERENCES

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. *Multimodal fusion for multimedia analysis: A survey*. Springer Multimedia Systems Journal, 16(6):345-379 (2010).

[2] M. Andersson, S. Ntalampiras, T. Ganchev, J. Rydell, J. Ahlberg, N. Fakotakis, *Fusion of acoustic and optical sensor data for automatic fight detection in urban environments*, Information Fusion (FUSION), 2010 13th Conference on , pp.1-8, 26-29 July 2010

[3] T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, *Audio-Visual fusion for detecting violent scenes in videos*. In Artificial Intelligence: Theories, Models and Applications, Vol. 6040 (2010), pp. 91-100.

[4] W. Zajdel, J.D. Krijnders, T. Andringa, D.M. Gavrila, *CASSANDRA: audio-video sensor fusion for aggression detection*, Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on , pp.200-205, 5-7 Sept. 2007

[5] Y. Wu, E.Y. Chang, B.L. Tseng – *Multimodal metadata fusion using causal strength*, MULTIMEDIA '05 Proceedings of the 13th annual ACM international conference on Multimedia, 2005.

[6] H. Xu and T.-S. Chua, *The fusion of audio-visual features and external knowledge for event detection in team sports video*, Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval MIR 04 (2004)

[7] N. Poh, J. Kittler - *A Unified Framework for Biometric Expert Fusion Incorporating Quality Measures*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2012, PG. 3-18.

[8] Z. Yang. *Multi-Modal Aggression Detection in Trains*. PhD thesis, Delft University of Technology, 2009.

[9] I. Lefter, L.J.M Rothkrantz, G.J. Burghouts, Z. Yang, P. Wiggers – *Addressing Multimodality in Overt Aggression Detection***.** TSD'11 Proceedings of the 14th international conference on Text, Speech and Dialogue, volume 6836 of Lecture Notes in Computer Science, page 25-32. Springer, (2011).

[10] W. W. Cohen: *Fast Effective Rule Induction*. In: Twelfth International Conference on Machine Learning, 115-123, 1995.