# Fusing Affective Dimensions and Audio-Visual Features from Segmented Video for Depression Recognition*

## INAOE-BUAP's Participation at AVEC'14 Challenge


Humberto Pérez Espinosa†
BUAP
Computer Science Department
Av. San Claudio y 14 sur
Puebla, 72000, Mexico
pesphum@gmail.com

Hugo Jair Escalante
INAOE
Computer Science Department
Luis Enrique Erro No. 1,
Tonantzintla,
Puebla, 72840, Mexico
hugojair@inaoep.mx

Luis Villaseñor-Pineda
INAOE
Computer Science Department
Luis Enrique Erro No. 1,
Tonantzintla,
Puebla, 72840, Mexico
villasen@inaoep.mx

Manuel Montes-y-Gómez
INAOE
Computer Science Department
Luis Enrique Erro No. 1,
Tonantzintla,
Puebla, 72840, Mexico
mmontesg@inaoep.mx

David Pinto-Avedaño
BUAP
Computer Science Department
Av. San Claudio y 14 sur
Puebla, 72000, Mexico
dpinto@cs.buap.mx

Veronica Reyes-Meza
UPAEP
Psychology Department
21 sur 1103
Puebla, 72410, Mexico
veronica.reyes@upaep.mx



## ABSTRACT

Depression is a disease that affects a considerable portion of the world population. Severe cases of depression interfere with the common live of patients, for those patients a strict monitoring is necessary in order to control the progress of the disease and to prevent undesired side effects. A way to keep track of patients with depression is by means of online monitoring via human-computer-interaction. The AVEC'14 challenge [13], aims at developing technology towards the online monitoring of depression patients. This paper describes an approach to depression recognition from audiovisual information in the context of the AVEC'14 challenge. The proposed method relies on an effective voice segmentation procedure, followed by segment-level feature extraction and aggregation. Finally, a meta-model is trained to fuse monomodal information. The main novel features of our proposal are that (1) we use affective dimensions for building depression recognition models; (2) we extract visual information from voice and silence segments separately; (3) we consolidate features and use a meta-model for fusion. The proposed methodology is evaluated, experimental results reveal the method is competitive.



## Keywords

Depressive disorder, Depression recognition, Multimodal information processing, Meta-classifier


## 1. INTRODUCTION

Depression is a disease that affects a considerable portion of the world population. According to the World Health Organization [9], up to 2012 there were at least 350 million people living with some form of depression (other sources report only over 121 million[1]). Severe cases of depression interfere with the common live of patients, in fact depression is the leading cause of disability in the world [9]. For those patients, a strict and almost permanent monitoring is necessary in order to control the progress of the disease and to prevent undesired side effects.

A way to keep track of patients with depression is by means of online monitoring via human computer interaction and affective computing technologies. The AVEC'14 challenge [13] is a first effort towards developing decision support tools that can help patients and therapists to keep track of the progress of the disease. In particular, the challenge focuses on the analysis of audiovisual information recorded from patients performing a predefined task. There are two tracks in the AVEC'14 challenge: the *Affect Recognition Sub-challenge* (ARS), which focuses on predicting continuous affective dimensions (Valence, Arousal and Dominance) and the *Depression Recognition Sub-challenge* (DRS), which


*This research was partially supported by the LACCIR programm under project id R1212LAC006.

†This author was supported by CONACYT throughout postdoctoral grant No. 49296-290807.




*AVEC'14*, November 7, 2014, Orlando, FL, USA.



http://dx.doi.org/10.1145/2661806.2661815.


[1]http://www.healthline.com/health/depression/statistics-infographic

directly focuses on depression analysis. This work focuses on the latter track of the AVEC'14 challenge.

The goal in the DRS is to develop methods that can automatically predict the value of a self-reported depression indicator from an audiovisual recording [13]. A system capable of making predictions with acceptable rates, would be very useful not only for monitoring[2] patients with a depression diagnostic, but also, to identify people who may have the disease and do not know it. The scenario considered in the challenge is appealing because it is a non-invasive procedure that does not rely on specialized equipment and can be applied massively.

However, although appealing, the DRS is very challenging in multiple ways: there is a small sample of training and development instances for building a predictive model; information was recorded with a standard webcam and under uncontrolled conditions; the users were not advised to maintain a fixed position/behaviour with respect to the webcam and/or microphone; the variable to be predicted has been indicated by the patients themselves; there is a wide variety of subjects; and even for some videos there is no audio information at all (i.e., the user did not produce a word during the recording).

In spite of these challenging conditions we propose a solution to the AVEC'14's DRS that aims at overcoming, to some extent, some of these limitations (see Figure 1). The proposed solution decomposes clips into segments (with and without voice) for feature extraction. Next, affective dimensions and audio-visual features are extracted from clip segments and latter aggregated to represent each clip. Monomodal predictive models were trained using this representation. Two prediction strategies were considered: on the one hand, we used the direct prediction of the best model, and on the other hand, we trained a meta-classifier on top of the predictions of all of the models. We evaluate the proposed methodology using training and development data and show that our best result is competitive with the baseline method supplied by the organizers [13].

A distinctive feature of our approach is that it relies on predictions of affective dimensions (arousal, valence, and dominance) *at the segment-level* for building a predictive model for depression recognition. Hence, providing evidence on the usefulness of affective dimensions on depression recognition. Another key aspect of our proposal is that we wanted to compliment audio information by incorporating visual information from silence segments. The underlying hypothesis was that visual features may be more helpful when the users are not vocalizing. The rest of the paper describes the components of our formulation and assess its performance in the context of the AVEC'14's DRS.

The paper is organized as follows. Next section briefly describes the considered scenario. Section 3 describes the methods we adopt for affective dimension prediction and clip segmentation. Section 4 describes the additional feature extraction process from audio and visual modalities. Section 5 introduces the fusion methodology and the overall approach to depression recognition. Section 6 presents experimental results that aim at evaluating the performance of the proposed solution. Section 7 discusses our main findings.

[2]One should note that, in the authors' opinion, this type of systems should be seen as support tools that can alleviate the load of work for therapists and at the same time be available to any potential depression-patient for facilitating a depression diagnostic. Therefore, in our view, the goal of this type of systems is not to replace therapists or clinical monitoring/analysis procedures, but equipping health systems with support tools that can increase their scope.

## 2. AVEC'14'S DRS SCENARIO

The scenario considered in AVEC'14's DRS challenge is as follows, see [13, 14] for further details. Two sets of labeled videos were provided, called training and development, additionally a set of unlabeled videos was also provided. Videos in the training and development partitions could be used to develop the predictive model, whereas the unlabeled data set was designated for the evaluation of the models.

Each of the data sets (i.e., training, development and test) contained videos taken from subjects with some degree of depression. Two types of videos were available per each subject *North Wind* and *Free Form* versions, the two types correspond to two different tasks the participants had to do in front of a computer equipped with webcam and microphone. There were available 50 videos from each type in each of the data sets, thus there were 100 videos per data set.

Training and development videos were manually labeled with a depression level indicator (for DRS) and the affect dimensions valence, arousal and dominance at frame level, 30 frames per second (for ARS). In the DRS, the labels for the videos correspond to the patient's Beck Depression Index-II (BDI-II), which was derived after participants filled a questionnaire [2]. The BDI-II is a numerical value between 0 and 63 (the largest BDI-II we found in training and development partitions was 45) that roughly indicates the degree of depression.

Regarding the affect dimensions, although they were supposed to be used for the ARS [13], in our approach these are used as attributes to build the depression recognition model (see Section 3). For training and development videos, the attributes were the ground truth labels, while for the test set, we built a model for affect dimension prediction from the ground truth labels and use the predictions of this model on the test data set as attributes.

Summarizing, the goal of AVEC'14's DRS is to generate a predictive model capable of associating test videos with the correct BDI-II label. Where the effectiveness of the predictive model is assessed by two main evaluation measures: the *Mean Absolute Error* (MAE) and the *Root Mean Squared Error* (RMS).

## 3. AFFECTIVE DIMENSIONS PREDICTION AND CLIP SEGMENTATION

As mentioned above, a distinctive feature of our proposal is the incorporation of affective dimensions as attributes to build a depression recognition model. The main question we wanted to answer with this idea was: *How strongly correlated are the affective dimensions to the depression indicator?* During the development phase we observed that affective dimensions could be better predicted on a segment-level basis (opposed to a frame/ms base), however, it was unclear what clip-segmentation method to use and what segment-size would be better option. Therefore, we implicitly aim to answer another related question: *What is the appropriate segment size to estimate more accurately valence, arousal and dominance?* This section elaborates on the answers for both questions.

## 3.1 Affective dimensions for DRS

In order to answer the first question, we calculated the Pearson correlation coefficient between the ground truth values of affective dimensions and the BDI-II value for training videos. The comparison was made at the segment level (see below the description of our segmentation technique); that is, using the averages of the affective dimensions at frame level. The goal of this analysis was to determine whether it would be beneficial to use affect predictions as attributes to estimate the depression indicator. Table 1 shows the Pearson correlation value between each affective dimension and the BDI-II for the training set.

**Table 1: Pearson correlation coefficient for training data. We report the correlation between affective dimensions and the BDI-II.**

| Primitive | Northwind | Freeform |
|---|---|---|
| Arousal | -0.45 | -0.32 |
| Dominance | -0.44 | -0.20 |
| Valence | -0.46 | -0.46 |
| Average | -0.45 | -0.32 |

It can be observed that, indeed, there is a significant degree of (negative) correlation between affective dimensions and depression indicator, mainly in the *North Wind* task. It is interesting that for this data type roughly the same correlation value was observed for the three dimensions. On the other hand, the correlation seems to be less strong for the *Free Form* videos.

Table 2 shows the Pearson correlation values between the affective dimensions. It can be observed that there exists some degree of correlation between affective dimensions and depression scores (slightly larger between Arousal and Dominance), which may explain the results from Table 1. However, the correlation was lower than expected.

**Table 2: Pearson correlation coefficient matrix for affective dimensions and depression labels.**

| Primitive | A | D | V |
|---|---|---|---|
| A | 1 | 0.64 | 0.58 |
| D | 0.64 | 1 | 0.58 |
| V | 0.58 | 0.58 | 1 |

Results from Tables 1 and 2 were encouraging and motivated us to study the suitability of affective dimensions for building depression recognition models. Still an open question here is whether affective dimensions are reliable enough for detecting depression. However, in subsequent experiments, additional evidence on the relevance of affective dimensions for automatic recognition of depression will be presented.

## 3.2 Clip segmentation

In order to answer the second question (regarding the adequate segment size to estimate affective dimensions), we tested two segmentation methods. The first one is the segmentation based on Voice Activity Detection provided by the organizers. Whereas the second one is a method for detection of sound and silence intervals based on a threshold of sound intensity as implemented in [3]. We found that by using the latter method and adding a maximum length restriction, the minimum length of voice intervals was of 0.5 seconds and the maximum one was of 2 seconds. This segmentation method provides shorter segments than those obtained with VAD segmentation, where segments may become longer than 40 seconds. Smaller segments are desired in our approach as we wanted to locally model affective and audiovisual information.

Additionally, we implemented a method to discriminate linguistic and non linguistic vocalizations like gasps and sobs. This is in order to focus the feature extraction process (of audio and affective features) to clip zones where the users are vocalizing, similar to the work in [1]. Our method is motivated by the idea that affect is expressed intensively in short episodes and that emotions could change rapidly [11]. The method uses two criteria to decide if a segment is considered as voice or no-voice[3]: a) there are syllables detected in the segment; and, b) the output of a classifier based on chroma features trained with voice and no-voice samples.

For segmentation, we first converted the provided mp4 audio recordings to WAV-PCM format. Then we applied the proposed segmentation method and extracted, from each segment, affective dimensions and audiovisual features.

The final set of features we consider for depression recognition is explained in Section 4, the rest of this section details the process for generating affective dimensions. For obtaining such affective values we built a regressor using as instances segments (represented by the acoustic features provided by the organizers [13]), where the labels for segments were obtained by averaging the ground truth values of valence, activation and dominance for frames within each segment. In order to compare the effectiveness of both segmentation approaches for predicting affective dimensions we conducted an experimental evaluation. For this comparison we used data from only the training partition. We generated regression models using Support Vector Machines [12], and we evaluated the prediction performance of the three affective dimensions. For the evaluation we used a 10-fold cross validation procedure (over training videos). We handled separately Freeform and Northwind clips. Table 3 shows the results of the comparison.

**Table 3: Pearson correlation for training data.**

| Task | Arousal | Dominance | Valence |
|---|---|---|---|
| **Provided VAD Segmentation** | | | |
| Freeform | 0.5060 | 0.4764 | 0.5045 |
| Northwind | 0.6312 | 0.5565 | 0.2858 |
| **Proposed Segmentation** | | | |
| Freeform | 0.6477 | 0.6680 | 0.3771 |
| Northwind | 0.4532 | 0.6430 | 0.5781 |

It can be seen from Table 3 that the correlation is stronger when using the proposed segmentation technique, this result holds for both types of videos, *North Wind* and *Free Form*. Please note that the correlation with the VAD approach is higher than that of our method in *Free Form* -

[3]Any sound that is not voice, which is different from silence segments.

*Valence* and *North Wind - Arousal*. Nevertheless, on average, our method outperforms significantly the baseline approach. Given these results, we decided to use our proposed segmentation method for our depression recognition system. The next section describes the set of features considered for building depression recognizers.

# 4. FEATURE EXTRACTION

In order to process clips we extracted descriptive features derived from both: the sequence of images and the audio signal extracted from the video. In the following we present the audiovisual attributes we considered. Please note that the attributes were extracted at a segment level using the method introduced in Section 3.2.

## 4.1 Audio based features at clip level

We designed a feature vector to represent/describe the behavior of subjects in the videos based on the audio signal analysis. We used regression models, as described in the previous section, to generate predictions for affective dimensions and use them as attributes at the segment level. Subsequently, we obtained the average of the predictions of each emotional dimension for all the segments of the clip. Finally, feature vectors are generated using these affective attributes per each clip besides other attributes as described below.

A total of 11 attributes per clip were calculated. These used attributes were:

1. Averaged dominance along clip
2. Averaged arousal along clip
3. Averaged valence along clip
4. Averaged speech rate along clip (number of detected syllables by [5] method /segment duration ).
5. Number of silence intervals greater than 10 seconds and less than 20 seconds.
6. Total time, in seconds, of silence intervals greater than 10 seconds and less than 20 second.
7. Number of silence intervals greater than 20 seconds
8. Total time, in seconds, of silence intervals greater than 20 seconds
9. Percentage of total voice time classified as neutral
10. Percentage of total voice time classified as happiness
11. Total duration of voice intervals

These 11 attributes were combined with the visual attributes described below to represent videos. Then predictive models were built upon this attributes and a fusion scheme was adopted for generating the final prediction.

## 4.2 Visual features

As video descriptors we considered a variety of general features that mainly comprise motion and velocity information. We used these general descriptors because we think they are not severely affected by the conditions in which video was recorded or by the position of subjects. What is more important, we hypothesize that motion and velocity information may reveal useful information about the depression status of subjects. Our hypothesis is funded in studies that suggest that subjects with depression may present psychomotor retardation, see e.g., [4]. Also, we visually inspected some training and development videos, and roughly found that, in fact, subjects with large BDI-II showed slower movements than those with small BDI-II.

We detected face and eyes bounding boxes by using the Viola-Jones detector [15]. After detection, the face was isolated from the whole image and features were extracted from the video segment containing the face. First, we consider the difference of final and initial positions of face and eyes within the video segment as well as the average, minimal and maximal coordinates of face and eyes (with these features we measure the range of movement/variation of face and eyes across the whole video). Also we consider the average velocity of face and eyes in $x$ and $y$ axis. Finally, we also extract the motion history image, motion static image and motion average image from the segment of video that contains the face only, we used implementations from [6]. In preliminary experimentation we compared the performance of our set of features and the baseline features provided by the organizers [13, 14]. Our experimental study revealed that our features were more discriminative than the baseline ones. Therefore, we decided to use our set of features.

# 5. PROPOSED METHOD

The proposed methodology is summarized as follows, see Figure 1. Training videos were segmented into voice and silence segments by considering audio information. Visual and audio features were extracted from each of the segmented videos. A feature selection procedure was applied independently for each modality (video, audio), segment type (audio, video-silence and video-voice) and for each task (*North Wind* and *Free Form*, see Section 2). Selected audio-based features were used to built a model for affective dimension prediction, whereas selected video-features were used directly for the recognition model. Each video is represented by the aggregation of the features extracted from all of its segments (consolidation process). A training set of attributes-labels pairs was generated with the video representation and the depression indicator the video. Monomodal predictors were trained, a fusion strategy was adopted to combine the predictions of the mono-modal models, and generate a single prediction for each sample. The rest of this section explains in detail some aspects of the proposed methodology.

## 5.1 Video segmentation

Videos were segmented into voice and silence segments by using the technique described in Section 3.2. The motivation to segment video into voice and silence segments was to extract features only from informative segments. In the case of audio, voice segments are the ones that might contain relevant information, whereas for video both, voice and silence segments, could be very helpful. In fact, a hypothesis of our work is that visual information extracted from silence segments may complement audio features (extracted from voice segments).

## 5.2 Feature extraction and consolidation

The features described in Section 4 were then extracted from each of the segments. After feature extraction, a feature selection process was conducted in order to reduce the dimensionality of data. We used the *relief* method in WEKA with default parameters [7] for this purpose.

After feature selection, a consolidation procedure is applied. This process aims at combining all of the information

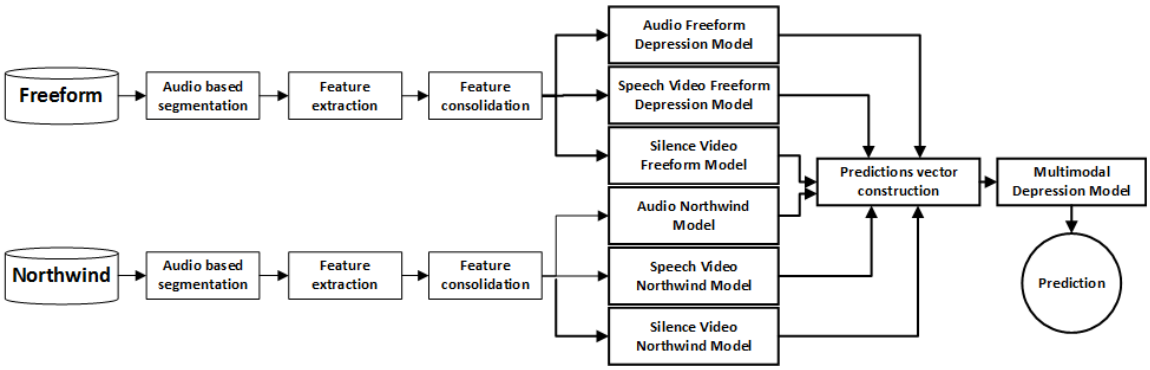


**Figure 1: General diagram of the proposed approach.**

from segments extracted from the same video, to represent the video. Specifically, to represent each video we take the average over the feature vectors of segments associated to the video. The idea of segmenting videos and then aggregating segment-attributes was based on the intuition that, as previously mentioned, features are extracted from segments of the video that are the most potentially useful. Hence, after the consolidation process each video is represented by a single highly-informative feature vector per information modality. Please note that we aggregate vectors from features extracted from the same type of feature-segment (i.e., audio-voice, video-voice and video-silence), see Figure 1.

## 5.3 Model construction

Once each video is represented by a vector of attributes, we proceed to build a predictive model. The model must be capable of associating feature vectors with the correct BDI-II of subjects. Although the BDI-II is a categorical variable, we treated the problem as one of regression. This is because of the number of categories and the few examples we had from each label to train a model (for some BDI-II values there are no samples at all). For building regressors we considered Gaussian processes [10] and support vector machines regressors [12], as implemented in WEKA [7], default parameters were considered for both methods. For each modality (audio-voice, video-voice, video-silence) and video type (*North Wind* and *Free Form*) we built a predictive model, see Figure 1. At this stage, we were able to make predictions by using any of the trained models.

Instead of making predictions directly from the predictive models, we can adopt an information fusion technique to combine the predictions from the six models. The intuitive idea is that different models that were trained on data from different sources, may be complimentary to each other. For combining the outputs of the mono-modal models we built another predictive model (i.e., a meta-classifier) that was trained on the outputs of the mono-modal predictions. This meta-model (a Gaussian process regressor was used), then combines the predictions of multiple mono-modal models, see Figure 1. In the next section we evaluate the performance of the proposed approach using training and development data, besides we report results in the test set for runs we selected to be submitted for evaluation in AVEC'14.

## 5.4 Alternative predictive models

It can be seen from Figure 1 that alternative ways for making predictions can be conceived by using the same architecture. Accordingly, and for the sake of comparison, we evaluate the performance of other variants, including direct prediction and majority voting.

Under direct prediction we consider the best individual model and use the predictions of this model for labeling test videos. In preliminary experimentation we found that the best individual model was the one based on audio-features only (see the Figure 1).

On the other hand, a majority voting strategy can be implemented to combine information from multiple segments, prior to the construction of the meta-model. That is, instead of performing the feature consolidation step, we can make predictions for each of the segments that form a video, and then, we can take the mode of predictions for segments within a video as attribute for the meta classifier.

# 6. EXPERIMENTS AND RESULTS

This section reports experimental results obtained with our proposed methodology in the AVEC'14 DRS challenge. First we evaluate the recognition performance when using only the affective dimensions. Then, we evaluate the performance of the direct-prediction strategy and then the performance of the meta-classifier technique. For the latter experiments we used training and development data. Finally we report the performance of our best runs in the test set.

## 6.1 Depression recognition with affective dimensions

Our proposed approach combines information from affective dimensions with audio and visual features. Because a novel component of our method is precisely the use of affective dimensions, it is worth analyzing the recognition performance when using affective dimensions only. Table 4 shows depression recognition results for both types of videos as well as for their combination. For this experiment we trained a support vector regressor in the training partition and made predictions for the development set. The reported performance, therefore, corresponds to the development set.

**Table 4: Results for depression scores predictions from affective dimension labels**

| Task | Correlation | MAE | RMS |
|---|---|---|---|
| Freeform | 0.4583 | 8.2976 | 11.2962 |
| Northwind | 0.5224 | 7.906 | 10.9192 |
| Both | 0.5224 | 7.906 | 10.9192 |

From Table 4 it can be seen that the performance of affective dimensions is quite competitive, we obtain performance that compares with the baseline. Slightly better results were obtained for the *North Wind* data set, in fact the combination of videos from both data types did not boosted the performance of the regressor.

Results from Table 4 are very interesting: they are initial evidence on the usefulness of affective dimensions for depression recognition; to the best of our knowledge, affective dimensions have not been used previously as features for depression recognition. Although the results are not as good as we wish, it is clear that we can develop other methods/representations that can take more advantage of affective dimensions for recognition. The next section shows that the combination of these features with additional information and under our proposed approach results in better recognition performance.

## 6.2 Direct prediction

This section reports the performance of the direct prediction strategy: making predictions with mono-modal models trained on consolidated features, see Figure 1. Table 5 shows the depression recognition performance obtained by the direct strategy when using each of the mono-modal techniques. For these experiments, the models were trained using the training data set and the performance of models was evaluated on development data sets.

For this particular experiment we used the Relief [8] feature selector in order to use only the best features from each modality. For the audio modality we used a support vector machine regressor, whereas a Gaussian process was used to generate the video based models.

**Table 5: Experimental results obtained by the direct prediction approach using consolidated attributes.**

| Modality | Correlation | MAE | RMS |
|---|---|---|---|
| **North Wind** | | | |
| Audio | 0.4811 | 8.902 | 10.6195 |
| Video Voice | 0.3156 | 9.4721 | 11.51 |
| Video Silence | 0.4573 | 9.6723 | 11.18 |
| Audio+Video* | 0.6026 | 7.7969 | 9.7873 |
| **Free Form** | | | |
| **Audio** | **0.6864** | **7.4895** | **8.9676** |
| Video Voice | 0.1146 | 8.64 | 10.4754 |
| Video Silence | 0.0614 | 8.7861 | 10.2169 |
| Audio+Video* | 0.6534 | 7.4723 | 9.0336 |

Audio+Video (*) means that audio features were combined with both Video Voice (VViddeo) and Video Silence (SVideo).

It can be seen from Table 5 that the performance of most of the models is competitive with the corresponding baseline results [13]. Better results were obtained for the *Free Form* data sets, although even the performance on *North Wind* was somewhat positive. Regarding the performance when considering different modalities, it can be seen that using audio alone and combining audio and video features yield the best performance. Using attributes extracted from video on silence segments alone did not result in competitive performance.

Comparing the best result from Table 5 and those from Table 4, it can be seen that the inclusion of audio features (in addition to the affective dimensions) resulted in slightly better performance in MAE, however, the difference in terms of RMSE is considerable. Hence, it seems that affective dimensions and the audio features from Section 4 are complimentary.

Table 6 shows the performance of the direct prediction strategy when using majority voting instead of feature consolidation to generate predictions.

**Table 6: Experimental results of direct prediction by majority vote approach.**

| Modality | Correlation | MAE | RMS |
|---|---|---|---|
| **North Wind** | | | |
| Audio | 0.43804 | 8.7660 | 10.800 |
| Video Voice | 0.16385 | 9.7447 | 11.832 |
| Video Silence | 0.38159 | 9.7692 | 11.419 |
| Audio+Video | 0.4678 | 9.1763 | 10.5641 |
| **Free Form** | | | |
| Audio | 0.34598 | 10.146 | 13.447 |
| Video Voice | 0.23876 | 8.6591 | 10.714 |
| **Video Silence** | **0.32435** | **8.4634** | **9.8414** |
| Audio+Video | 0.3759 | 9.1512 | 11.0124 |

From Table 6 it can be seen that, in general, the feature consolidation approach (see Table 5) outperform the majority voting strategy (Table 6). Thus, it seems that the feature consolidation approach can be more beneficial for the meta-classifier fusion strategy. It is interesting that for the majority voting strategy the best result was obtained when using the model from video-silence.

## 6.3 Meta model

Once we have analized the performance of mono-modal predictors, we assess the performance of the proposed meta-classifier based fusion approach. Table 8 shows the results of this experiment. We show the performance of the meta-classifier approach when this model is trained from the outputs of predictors trained on consolidated features, vs. when considering the majority vote outputs. For this experiment, the models were trained on the training data set and the performance was evaluated on the development data set.

**Table 7: Experimental results of meta-classifier.**

| Modality | Correlation | MAE | RMS |
|---|---|---|---|
| **Feature consolidation** | | | |
| Audio+Video | 0.7261 | 6.7862 | 8.3058 |
| **Majority vote approach** | | | |
| Audio+Video | 0.5209 | 7.9641 | 10.1376 |

It can be seen from this table that, although both methods offer competitive performance, better results were obtained with the feature consolidation formulation. The performance of the meta-classifier trained on consolidated features outperforms the results reported by the organizers in [13].

## 6.4 Results on test data

From the results obtained in the previous two sections, we can select the configurations of our method that are to be evaluated on test data (recall that the evaluation on test data was done by the organizers, see [13]). Specifically, we decided to submit, on the one hand, the best mono-modal direct prediction model and on the other hand the best configuration obtained with the meta-classifier. Table 8 shows the test set performance of our best models.

**Table 8: Results on submitted predictions**

| Modality | MAE | RMS |
|---|---|---|
| **Direct Prediction** | | |
| Audio Freeform | 9.3539 | 11.9165 |
| **Meta-classifier** | | |
| **Audio+VVideo+SVideo** | **8.9910** | **10.8239** |

From Table 8, it can be seen that (in agreement with our experimental study) better results were obtained by the meta-classifier. As before, the performance of this variant is competitive with the baseline approach described in [13].

# 7. CONCLUSIONS

This paper introduced a novel approach to depression recognition from videos. The distinctive features of our proposal are a voice/silence segmentation process, the use of affective dimensions as attributes, a feature consolidation procedure and a fusion scheme based on a meta-classifier. We assess the performance of different aspects of the proposed approach and compare it with alternative techniques. From the experimental analysis we can conclude:

- Our experimental study reveals that using affective dimensions as attributes for depression recognition is a promising and fruitful approach. In addition to the performance assessment, we found that when mixing affective dimensions with more than 2000 acoustic features and applying a feature selector, affective dimensions always appeared within the top-5 most discriminant features.
- We found that the proposed clip-segmentation approach performs better than the one provided by the organizers. This is due to the fact that smaller voice segments are generated, and that we applied a postprocessing to remove sounds other than voice from segments.
- The direct prediction strategy proved to be very effective when compared to the other variants. Specifically, a mono-modal audio-based model, is quite competitive with our multi-modal approach.
- The meta-classifier based fusion-scheme proved to be very helpful for depression recognition. In particular, the meta-classifier over mono-modal models trained on consolidated features. This result also provides evidence on the usefulness of feature aggregation.

# 8. REFERENCES


[1] A. Batliner, D. Seppi, S. Steidl, and Björn Schuller. Segmenting into adequate units for automatic recognition ofemotion-related episodes: A speech-based approach. *Advances in Human-Computer Interaction*, 2010(Article ID 782802):15 pages, 2010.

[2] A. Beck, R. Steer, R. Ball, and W. Ranieri. Comparison of Beck depression inventories -ia and -ii in psychiatric outpatients. *Journal of Personality Assessment*, 67(3):588–597, 1996.

[3] Paul Boersma and David Weenink. Praat: doing phonetics by computer. July 2010.

[4] J. S. Buyukdura, S. M. McClintock, and P. E. Croarkin. Psychomotor retardation in depression: Biological underpinnings, measurement, and treatment. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 35(2):395–409, 2011.

[5] NivjaH. de Jong and Ton Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, 2009.

[6] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner. Results and analysis of the chalearn gesture challenge 2012. In *WDIA: Advances in Depth Image Analysis and Applications*, volume 7854 of *LNCS*, pages 186–204. Springer, 2013.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[8] Kenji Kira and Larry A. Rendell. A practical approach to feature selection. pages 249–256, 1992.

[9] World Health Organization. Depression - a hidden burden. who flyer. http://www.who.int/topics/depression/en/, Link verified on July 15, 2014 2012.

[10] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[11] Klaus R. Scherer. *Psychological Models of Emotion*, chapter 6, pages 137–162. Oxford University Press, 2000.

[12] B. Schölkopf and A. J. Smola. *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[13] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. AVEC 2014 - 3D dimensional affect and depression recognition challenge. In *Proc. of the 4th ACM international workshop on Audio/visual emotion challenge*, 2014.

[14] M. F. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. AVEC 2013 - the continuous audio/visual emotion and depression recognition challenge. In 3-10, editor, *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, 2013.

[15] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Vision and Pattern Recognition Conference*, volume 1, pages 511–518, 2001.