Analysis of different affective state multimodal recognition approaches with missing data-oriented to virtual learning environments

Camilo Salazar, Edwin Montoya-Múnera, Jose Aguilar

*Corresponding author: CEMISID, Universidad de Los Andes, Mérida, Venezuela.
E-mail address: aguilar@ula.ve (J. Aguilar).

https://doi.org/10.1016/j.heliyon.2021.e07253
Received 23 February 2021; Received in revised form 31 March 2021; Accepted 3 June 2021

1. Introduction

Virtual education has been gaining great popularity around the world due to cheapening of communication devices, globalization of the use of the internet, advances in connectivity and universal access to educational resources. In addition, governments are every day more interested in virtual education for the lower costs compared with classroom education, accessibility, ease of schedules, among other aspects. Additionally, recent advances in communication technology make virtual education available to all countries.

According to [1, p. 1] “There is empirical evidence in the psychological literature that affective states affect cognitive processes like memorizing and decision making.”. In general, the relationship between emotional states and the learning process has been widely studied in the psychological field, and several scientific studies link emotions to education [2, 3, 4, 5, 6, 7]. According to these studies, emotions affect the way people learn, they are very correlated with essential elements of students’ self-regulated learning, such as interest, motivation, and strategies of learning [8].

Recognizing the affective states in virtual education has particularities that must be taken into account for an accurate assessment, mainly, the sources of information for recognizing affective states should not be obtrusive or invasive for not perturbing the learning process. For example, many works use electroencephalograms for recognizing emotions in people [9, 10, 11, 12], which cannot be used in an educational context because this sensor can be annoying for students, disturbing the natural learning process. Sources of information must allow the free development of the learning process, and these could be cameras, microphones, text from chats, reviews, etc. Another important consideration is that in virtual learning environments not all the sources of information (also called modalities) are available at every moment. In a virtual learning environment, a student is not all the time talking or writing. Due to this, the system for recognizing the affective states in virtual education must consider the availability of modalities at all times.

During the recognition of affective states, the more information available the more accurate the recognizing can be. A single instant can not reflect all the context of a situation, periods of time must be considered rather than single moments. In a single instant, an image can suggest that a person is asleep, but in a video, a sequence of images can show that it was only a blink. This work aims to recognize affective states in periods of time instead than on specific moments, using infor-
motion from all the sequence of moments that compose the period of time.

No one modality is good enough for recognizing an emotion in an arousal and valence emotional model, some are very accurate for arousal while others are more for valence [13], which also depends on the situation. Mixing multiple modalities help strengthen weaknesses among themselves, and therefore, having a more robust affective state recognizer. Multimodality offers more points of view for taking a decision. In our proposal, we consider three modalities: audio, video and text, which are not-intrusive or invasive to the learning process and can be captured from affordable devices, such as integrated cameras or microphones.

The main contribution of this work is to propose and compare approaches for the recognition of emotional states in terms of arousal and valence dimensions from various non-intrusive and affordable modalities, using only the available modalities at each specific moment, oriented to the context of the virtual education.

The rest of the paper is structured as follows: Section 2 exposes the theoretical framework of this work. In Section 3 related works are reviewed and compared with our approach. The proposed approach is described in Section 4. Experiments and implementation details are given in Section 5. In Section 6 results and analysis are provided. Finally, in Section 7 some conclusions, discussions and future works are presented.

2. Theoretical framework

2.1. Multimodal data fusion

Modalities represent different sources of information, which can suggest different information, different points of view. To obtain a global perspective, the system must have all the available information. In multimodal emotion recognition, different modalities can predict different affective states. Each modality has advantages and weaknesses: audio modality is better than video modality in poor lighting conditions, and text modality can be better than audio modality for predicting valence dimension under certain circumstances. Since modalities can have flaws in specific situations, information from other modalities must be considered for improving recognition in those scenarios. The idea behind multimodal fusion is to strengthen the weaknesses of each modality using the joint information to obtain a more robust system.

For multimodal data fusion, there are two main approaches: feature level fusion and decision level fusion. Fig. 1 graphically shows the differences between them. The first consists of fusing the characteristics extracted from the different modalities (several works use concatenation as a fusion technique), and then, using a model for predicting the target from the fused characteristics. The second requires unimodal models before fusion for making separate predictions of each modality, and then, fusing these predictions using an ensemble model to obtain a final prediction.

2.2. Models for representing emotions

The representation of emotions has been studied during the last century by different psychologists. There are two groups of models for representing emotions: categorical and continuous models. Categorical models use discrete values for representing emotions, and these models are more intuitive for humans. For example, Ekman proposed a model with 6 discrete emotions [14]: Anger, Disgust, Fear, Happiness, Sadness and Surprise. In Table 1 various categorical models are exposed.

The second group consists of continuous models with one or various dimensions for representing affective states. Empirical evidence suggests that continuous models are more accurate with respect to discrete models [13]. In this work, the arousal-valence dimensional model is used, the valence dimension for the pleasure level and the arousal dimension for the excitement of the person. In Fig. 2, an example of a map of some emotional terms for an arousal-valence continuous model is presented (various psychologists propose different mappings).

3. Related works

Multimodal emotion recognition has several results recently. The authors of [17] make an amazing job of reviewing unimodal and multimodal approaches for emotion recognition for the modalities of audio, video and text. The paper [18] presents a review where most of the papers work with categorical models for the representation of the emotions, and just a few utilize continuous dimensions.

The work in [19] proposes a two steps approach for recognizing emotions in educational contexts, the first step indicates if in a period of time there is the presence of an educative relevant emotion, and if there is, the second step recognizes that emotion. They use video and physiological signals (heart rate, breath, skin conductance and temperature) as modalities. These last modalities are obtrusive for the learning process due to the devices required for the measure, additionally, those devices are not affordable for all students of virtual education. Salmeron also experimented with low intrusive modalities for affect recognition in [20], using interactions with mouse and keyboard from students in order to predict emotions and valence states in SAM [21] scale, but they do not consider the availability of these modalities at every moment. The investigation published in [22] uses the audio modality for detecting emotions for classifying positive and negative emotional clips. They extract a set of features based on the feature set proposed in the Inter-speech challenge [23], and classify the audio signal in categories using a linear support vector machine (SVM) model. They reported 75.51% of accuracy. The paper [24] extracted features from audio, visual and text modalities, in order to recognize categorical emotions. They used different datasets as training set and testing set for visual and textual modalities, but the audio was trained and tested with the same dataset. They concatenated the characteristics from the different modalities, and then used SVM as a classifier. They reported 87.95% of average accuracy. In [25], features for two modalities (video and text) were extracted using neural networks. They also used the Interspeech feature set for audio features. Then, they concatenated the features and used a recurrent multiple kernel learner as a classifier. They reported 76.85% of average accuracy for the IEMOCAP dataset [26]. These three works performed very well for unimodal and multimodal emotion recognition in the categorical model, but they require that the three modalities are available all time, which is not always the case in real conditions.

The authors of [27] use video modality (head movements) for predicting values over five dimensions (arousal, expectation, intensity,
power and valence). They use optical flow on faces for detecting head movements, then use a Support Vector Regression (SVR) as a model for predicting values for the five dimensions. They reported 0.094 MSE (mean squared error) on the SEMAINE dataset. The investigation [28] claims that including events such as smiles, head shakes or laughter can improve the prediction of values over the affective dimensions. They utilize a string-based method for fusing the features, used SVR as predictor and reported 0.19 MLE (mean linear error). The work [29] uses audio, visual and text modalities on the Audio/Visual Emotion challenge (AVEC) 2017. They present good results, close to 0.09 RMSE (root mean squared error), for each dimension. However, data do not reflect real conditions in virtual learning environments, they do not deal with missing data, and models used for extracting features and predicting affective states are complex neural networks that have high requirements of computational resources and time for predicting, which is not the best option for scalability.

A similar work to our approach is proposed in [30] that also deals with missing data for multimodal emotion recognition, which is a different problem to other approaches for the problem of the missing labels. They exposed various ways for dealing with missing data at the fusion level, and used one for the experiments. However, even though they use the arousal-valence dimensional model (a continuous model), they predict quadrants and not continuous values in that space. In this way, their approach is considered as a categorical model for representing emotions, which the precision can be improved if a continuous model is used, as it is claimed in [13]. They reported 55% of accuracy on the CALLAS expressive corpus [31], predicting one of the four quadrants in the arousal-valence dimensional model (see Fig. 2). Particularly, the work presented in [30] compares various approaches for dealing with missing data, being the majority voting approach the best one. This is an approach only possible for categorical classifications and not for continuous models. For that reason, it is not possible a direct comparison between this work and ours (our target is an emotional continuous representation). A similar approach for continuous models is averaging the outputs, strategy, which is used for our “base model” (explained in the “our approach” section).

Table 2 presents a comparison between these works and our approach based on the following criteria:

(A) More than 1 modality.
(B) The target is the continuous bi-dimensional space of arousal and valence.
(C) Uses non-intrusive and affordable modalities.
(D) Works with missing data.
(E) Oriented to virtual education.

Several recent works meet criterion A. Multimodal approaches have been shown to have advantages over unimodal approaches. On the
other hand, the recognition of discrete emotions is a traditional approach that is still very used, especially for works that focus on specific affective aspects and not on the whole emotional spectrum (see criterion B). However, nowadays works focusing on continuous models starting to be much more frequent. Criterion C is related to non-intrusive modalities. As [17] shows, several works use non-intrusive modalities. Criterion D refers to dealing with the availability of input data modalities at all times. Missing data is the most frequent case in virtual learning environments, the learner is not writing or speaking all time, making these two modalities available only at certain moments. Finally, criterion E is related to the context of application (virtual education). Few works for multimodal emotion recognition focusing in this field have been published, compared with other areas. In our work, this orientation is implicit in the dataset used, which is the most similar to the real data in a virtual learning environment.

A big difference between those works and our approach is considering the availability of modalities. In our work, we deal with the availability of modalities at every moment for reflecting the reality in virtual learning environments. During the search, it was not found a reason why not considering missing data for multimodal emotion recognition, it is not an exclusive problem of the virtual educational field. Also, in the literature, there are works for the problem of missing labels, as is the case of [32], but which is different from the problem solved in this paper about missing modalities in input data.

According to the literature review [18], our approach is the only one that includes the next characteristics: uses non-intrusive and affordable modalities (audio, video and text), deals with missing data (missing modalities) at all time in order to reflect the reality in the virtual education, and represents emotions using a continuous model for improving representation precision.

### 4. Our approach

The main objective of this work is the creation of a system for multimodal recognition of affective state, representing emotions in the continuous model of arousal and valence, using non-intrusive and affordable modalities, dealing with the availability of the modalities at every moment, reflecting the real conditions in a virtual learning environment. The proposed general procedure for affective state recognition from audio, face and text modalities consists of 4 main steps: feature extraction, feature selection, fusion and recognition, as is shown in Fig. 3. This general procedure is based on the common practices observed in the literature [24, 25, 27, 29, 33, 34, 35].

The general idea is to extract features from each modality, select only those features relevant to our objective, fuse the data from the three modalities (audio, text and video), and then recognize the affective state based on the multimodal data. The main contributions of this work are inside the purple steps in Fig. 3. Each step of the general procedure is defined below.

#### 4.1. Feature extraction

In this subsection, the feature extraction process of the general procedure is explained. From each modality, relevant features are independently extracted. As the objective of our work is not to investigate the extraction of unimodal characteristics or an unimodal affective state recognition, but rather the fusion of multiple modalities, the unimodal approaches were adapted from related works that showed good results. In concrete, the approaches presented in [24, 25] are used as the base for our feature extraction process. The specific implementation details are given in the experiment section.

##### 4.1.1. Audio

For the audio modality, the features from the Interspeech Compare 2013 feature set are selected, as proposed in [25]. This set consists of a total of 6373 features, including low-level descriptors and statistical metrics of them, like mean and standard deviation. This standardized feature set contains different features that include the “most common and at the same time promising feature covering prosodic, spectral and voice quality features” [23]. Some of them are the root mean square frame energy, the pitch frequency, and the harmonics-to-noise ratio. Also, it includes statistical descriptive metrics like the arithmetic mean, standard deviation, kurtosis, skewness, minimum and maximum values, among others.

##### 4.1.2. Video

For the video modality, 68 facial landmarks for every 3rd frame of each video are extracted using the dlib frontal face detector and face predictor [36]. There are better face detectors that work with no frontal faces and more accurate landmarks predictors that work better with overlapping objects or with bad lighting conditions, but this model works fine and requires little time, a very precious feature for scalability in virtual learning environments. In addition, the distances of each landmark to each other normalized by the height of the face detected are used as features for each frame. As a result, 2278 features per frame were extracted. For summarizing, the features of the sequence of frames that represents each video are averaged resulting in a total of 2278 features for each video.

##### 4.1.3. Text

For the text modality, the Senticnet [37] database was used as a knowledge base for extracting the affective features from text, specifically, the pleasantness, attention, sensitivity, aptitude, polarity, primary mood and secondary mood features were extracted for each word. Additionally, the AffectiveSpace resource [38] was used for extracting 100-dimensional vectors from each word. As a result, 105 continuous features and 2 categorical features were extracted from each word. Finally, the percentiles 0, 25, 50, 75 and 100 were used for continuous features and the sum of occurrences of each category for categorical features were used, resulting in 541 features per text: 5 percentiles for each continuous feature: 105 × 5 = 525, and 8 for each categorical feature because there are 8 classes: 2^8 = 16. Thus, 525 + 16 = 541.

#### 4.2. Feature selection

The feature selection process is guided by three filters, the macro-algorithm of this process can be seen in Fig. 4. The first filter removes features that present low variance, these features are almost constants with low information (see lines 1-3). The second filter eliminates features that can be represented by others in the set. This degree of explanation is calculated using the variance inflation factor (VIF), which is defined as (see lines 4-13 in the algorithm):

\[
VIF_j = \frac{1}{1 - R^2_{ij}}
\]

| Table 2. Comparison with related works. |
|-------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Work                          | A               | B               | C               | D               | E               |
| [19] X                        |                 |                 |                 |                 |                 |
| [20] X X X                    |                 |                 |                 |                 |                 |
| [22] X                        |                 |                 |                 |                 |                 |
| [24] X                        |                 |                 |                 |                 |                 |
| [25] X X                      |                 |                 |                 |                 |                 |
| [27] X X                      |                 |                 |                 |                 |                 |
| [28] X X X                    |                 |                 |                 |                 |                 |
| [29] X X                      |                 |                 |                 |                 |                 |
| [30] X X X                    |                 |                 |                 |                 |                 |
| Our work                      | X X X X         |                 |                 |                 |                 |

C. Salazar, E. Montoya-Muñera and J. Aguilar
Heliyon 7 (2021) e07253
\( R^2_{x_j} \) is the R-squared metric of a model that predicts the variable \( j \) from the other variables in the set. The VIF filter prevents from multicollinearity of features. VIF value is calculated for each feature, and if the maximum VIF is higher than a threshold, then the feature with maximum VIF is eliminated and the VIF calculation process starts again. The third and last filter prunes features by their importance for predicting the target. Models as decision trees provide information on the importance of features. Best features can be selected by adjusting a threshold for this importance (this filter is explained in lines 14-19). A model that provides this information is fitted, in this case, random forest, then all variables below an established threshold are removed from the set.

### 4.3. Fusion and recognition

The fusion and recognition steps depend on the fusion approach, and they can be carried out together or separated. As mentioned in the theoretical framework section, there are two strategies for data fusion: feature level fusion and decision level fusion. Decision level fusion requires unimodal models for making predictions before fusion, feature level fusion does not. This section proposes several fusion and recognition approaches, and is divided into two parts, the definition of some concepts required by our approaches and the presentation of our approaches.

#### 4.3.1. Definition of preliminary concepts

In this section are defined what is a unimodal model and a base model, concepts that will be used by our approaches.

**Unimodal models** The decision level fusion strategy needs the definition of unimodal models. They are required by the approaches that use a decision level fusion strategy to make the fusion of modalities. Each one of them receives as input the features extracted from its modality, and outputs values for arousal and valence dimensions. Since the target is two continuous values (one for arousal and one for valence), metrics selected for evaluation and comparison are \( R^2 \) and the root mean squared error (RMSE) scores. RMSE is defined as:

\[
\sqrt{\frac{1}{n} \sum (y_j - \hat{y}_j)^2}
\]
The $R^2$ score is defined as:

$$1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Where $\hat{y}$ is the predictions, $y$ the true values, $\bar{y}$ is the mean of $y$, and $n$ the number of samples.

$R^2$ is limited to one at the top, a higher value is better. RMSE is limited to 0 at the bottom, a lower value is better. Intuitive interpretations of the metrics are: $R^2$ is how good is the model compared to the model that always predicts the mean of the true values. RMSE is the mean error in the original units.

Additionally, the relative error of standard deviation was added as a metric for measuring how similar the dispersion of the predictions is to the dispersion of the labels. Like any other error, the lower the better. Relative error is defined as:

$$relative\_error = \frac{predicted\_value - true\_value}{true\_value}$$

As our target is bidimensional, the euclidean norm is used and its formula is as follows:

$$SRE = \frac{||(\hat{\sigma}_{aw}, \hat{\sigma}_{val}) - (\sigma_{aw}, \sigma_{val})||_2}{||(\sigma_{aw}, \sigma_{val})||_2}$$

Where SRE is the relative error of standard deviation, $\sigma_{aw}$ is the standard deviation of the arousal labels, $\sigma_{val}$ is the standard deviation of the valence labels, $\hat{\sigma}_{aw}$ is the standard deviation of the arousal predictions, $\hat{\sigma}_{val}$ is the standard deviation of the valence predictions.

**Base model** In this paper, three approaches for multimodal recognition dealing with missing data are proposed, and a base model is implemented in order to test the results. The base model is based on a decision level fusion, so it requires a model for each modality before fusion. In addition, it consists of aggregating the results of the three unimodal models, averaging each dimension. As average can be performed over one, two or three values, it can deal with the variability number of available modalities at all time. Also, this average is weighted by the calculated performance of each unimodal model, in order to give more relevance to better performing models. The prediction of this model is defined as:

$$prediction = \sum_{j=1}^{n} (aw_{ij}, val_{ij}) \times W_j$$

Where $n_j$ is the number of available modalities at moment $i$ ($n_j$ ranges from 1 to 3), $aw_{ij}$ is the predicted arousal for the moment $i$ and modality $j$, $val_{ij}$ is the predicted valence for moment $i$ and modality $j$, and $W_j$ is the weight given to modality $j$ according to its performance. These weights are calculated as:

$$W_j = \frac{1 - error_j}{\sum_{k=1}^{k} (1 - error_k)}$$

Where error is a metric of the error of the model, and $k$ iterates over the modalities available at moment $i$. In other words, the denominator is the sum of 1-errors of models from modalities available at moment $i$. These weights are pre-computed before prediction calculation and are static. The output of this model is a two-dimensional point that indicates the affective state in the arousal-valence dimensional model (see Fig. 2).

**4.3.2. Our fusion and recognition approaches**

**4.3.2.1. First approach: feature-level fusion with zero padding** The first approach uses feature level fusion and consists of concatenating the features extracted from each modality, filling with zeros the missing data, and using a model for predicting arousal and valence values. This approach intends to replicate the idea of dropout, widely used in the neural networks field. Fig. 5 shows the process when video modality is not available.

**4.3.2.2. Second approach: decision level fusion with zero padding** Since this approach is based on decision level fusion, it requires unimodal models for obtaining the predictions that are going to be fused. This approach is very similar to the previous one, but consists of filling with zeros (padding) the predictions that are not available, then concatenating these predictions, and using a model for generating a final prediction. Fig. 6 shows the process when video modality is not available.

**4.3.2.3. Third approach: decision level fusion with dynamic input** This approach consists of using models that allow inputs of varying length, in this way, there is no need to fill the input, but rather it can be passed regardless of its length. Specifically, recurrent neural networks are used, which through weight sharing allow sequences of any length to be input. These types of networks are unrolled according to the length of the input sequence, having a memory state that allows the next element in the sequence to know information about the past element(s), and thus generate a prediction with the information of the last element. This memory state stores information of all the previous elements. Fig. 7 shows the process when video modality is not available.

**5. Setup of the experiments**

This section describes experiments performed and the implementation details of each step of our general procedure.

**5.1. Feature extraction**

The approaches presented in [24, 25] are taken as bases, and their implementations are adapted for our specific purposes. These works were selected due to their good results for the recognition of the categorical emotions. These works propose Enterface and IEMOCAP as datasets, however, in our case different datasets were used for each modality (see artificial dataset subsection).

**5.1.1. Audio**

For this modality, the work presented in [25] is taken as a base. As proposed in this work, IEMOCAP was used as the dataset, Interspeech as the feature set, and OpenSmile as the software for extracting the features.

**5.1.2. Video**

The work presented in [24] proposes to use 66 landmarks, we instead use 68 landmarks for ease of implementation and the difference is not significant. We also used the distances between every pair of these landmarks as features, resulting in 2278 features per face. Additionally, we normalized these distances by the height of the detected face. With this modification, it is sought that the characteristics are comparable between different faces regardless of the distance they are from the camera, since the closer they are, the larger the recognized face will be, and therefore, the greater the distances will be. The neutral face preprocessing that they propose was not used. The dataset used for this modality is hci-tagging [39]. For summarizing features from each frame of a video (a video is a sequence of frames), the average of the features is used, resulting in 2278 features per video. Other approaches for summarizing were analyzed: summation, quartiles (0,25,50,75,100), max and min. The average was the one with the best performance, but summation and quartiles have similar performances to average.

**5.1.3. Text**

Authors of [24] propose using EmoSenticSpace [40] as a knowledge base, however, this resource is not available, and instead, we used AffectSpace [38] that also provides 100-dimensional vectors of terms describing the affective information. They propose using the polarity score provided by Senticnet [37] as an additional feature, instead we
used polarity and the other six features that Senticnet provides: pleas-
antness, attention, sensitivity, aptitude, primary mood and secondary mood, as explained in the previous section. The negation approach that they proposed was not used. The dataset used for this modality is arousal-valence Facebook Posts [41].

5.2. Feature selection

The feature selection process consists of three sequential filters: low variance, high VIF and low importance. For the first filter, the traditional variance was used and the threshold selected is a minimum of \(1 \times 10^{-4}\). For VIF, linear regression models were used as predictors, and a maximum VIF of 10 was used to admit for any feature. For the feature importance filter, random forest models were built for extracting the importance, and the minimum importance accepted was \(1 \times 10^{-3}\). Only those features that meet these requirements were kept.

The feature selection process is performed for each of the two dimensions (arousal and valence), and each modality (3 modalities). The total features before the selection process are: for audio: 6373, for video: 2278, and for text: 541.

5.2.1. Audio

For audio modality, the first filter removed 621 features that presented variance lower than \(1 \times 10^{-4}\). The second filter is expensive in time, especially for audio modality due to the big quantity of features. For each one, it is necessary to build a model for calculating its VIF value, then at first VIF step 5752 models must be calculated, in the second step 5751, and so on. Even thinking this process is executed just one time, it is very time-consuming. For that reason, a heuristic was used for speeding up the procedure: eliminate 1000 features by step instead of 1 (as it is done is traditional VIF filtering) until a limit, then 500 by step, then 250 by step, and so on. At the end of the VIF filter, only 2752 features were considered as not highly correlated. The final filter was only passed by 70 features for arousal and 65 for valence.

5.2.2. Video

The first filter did not eliminate variables for video modality, all features have a considerable variance. VIF filter showed that facial features are very highly correlated, therefore only 21 passed this filter, 2257 were removed for being highly correlated. As few features passed the VIF filter, feature importance is high due to the low number of variables, all the 21 features passed the third filter for both dimensions.

5.2.3. Text

As for video modality, the variance filter for text modality eliminated no features evidencing acceptable variance for every descriptor. 366 features were selected after the VIF filter, indicating that 175 features are highly correlated with the others in the set, thus were removed. 223 were considered relevant for predicting arousal and 200 for valence.

The feature selection process is very expensive in time, especially because of the calculation of VIF values for a large set of features. Despite that, as it is executed just one time, it does not represent a big problem.

5.3. Unimodal models

Since this model is not part of our main contribution, no in-depth research or experimentation was done. Techniques compared for every modality are: Random forest (RF), support vector machines (SVM) and Partial Least squares regression (PLS). RF and PLS, both allow multiple output as needed in this work, on the other hand, SVM does not sup-
port multiple output. Hence, two models were trained, one for arousal and one for valence, then the predictions are concatenated and metrics calculated. Some bias were found on training data, it was eliminated by removing high-density points in data: truncating each point in data to maximum one or two samples, depending on the modality. Labels closer than 0.05 in the euclidian distance are considered as the same point, and only one or two samples are accepted per point for reducing high-density points in the dataset that included bias in the data.

RF performed very well in training data, but had problems for generalizing, predictions in the test set presented low variance and were biased to the center of the space. On the other hand, PLS performed badly compared to RF in the train set, however, it presented better generalization with the test set. PLS was chosen as the model for the three modalities.

Text modality showed very bad results for the arousal dimension, which impoverishes the results of the model. It gives clues that recognizing arousal from social media posts is a very difficult task, and textual data may not be very informative about this specific affective dimension.

5.4. Artificial dataset

There was a great difficulty finding datasets that meet four requirements: 1) contain the three modalities (Audio, Video and text), 2) annotated in the arousal-valence space, 3) labels cover the majority of the arousal-valence space, 4) data in the English language. These four requirements do not assure good quality of data, but they are the minimal characteristics required to test our approaches. No dataset with the four characteristics was found, some of the explored datasets were: IEMOCAP, Enterface [42], Mosel [43], SEMAINE [44], MOSI [45], MOUD [46], MELD [47], Humaine [48, 49], CK [50], CK+ [51], DEAP [52], AMIGOS [53, 54], AFEW-VA [13]. Thus, it was decided to look for unimodal datasets that meet these requirements and as far as possible that they reflect the real conditions of virtual education, with the objective of fusing them for creating an artificial dataset that have the required features.

For the audio modality, IEMOCAP [26] was selected because it is a conversational dataset similar to circumstances in virtual learning environments, where learners use their voice primarily to conversationally communicating with others. Also, the data acceptably cover the space, as is shown in Fig. 8.

For the video modality, mahnob hci-tagging dataset [39] was chosen because in the dataset various participants were asked to watch movie fragments, and after watching the fragments, annotate their affective states using the self-assessment manikin. This is close to reality in virtual learning environments, where learners interact with content and obtain an affective state. In addition, data covers very well the space, as can be observed in Fig. 8. The disadvantage of this dataset is the number of samples: only 513.

For the text modality, it was difficult to find a dataset that contains arousal and valence annotations. Many datasets contain only valence annotations because this problem has been much more attacked in NLP for product reviews, marketing, etc. To the best of our knowledge, the more similar dataset to the real settings in virtual learning environments that is annotated in terms of arousal and valence is the arousal valence Facebook posts [41]. An advantage of this dataset is that it covers decently the space, as is shown in Fig. 8.

Fig. 8 shows that a dataset that meets the four conditions previously established can be built from these three datasets. For dealing with multimodality, datasets were fused using the following idea: if two samples from different datasets are assigned the same point in the arousal-valence dimensional model, those two samples express the same affective state, therefore, they can be joined in one sample with the same assigned value in the space.

Thus, the fused samples of the resultant artificial dataset are not related to a single individual; however, since the samples of the three modalities express the same affective state, they can be fused.

In practice, the process is the following: for each point in one modality, if there exists one point in each of the other two modalities that are closer than 0.05 (using Euclidean distance) to this original point, they are considered as a sample for the fused dataset. In other words, if the three points from different modalities are closer than 0.05 from each other, the samples are fused and considered as a sample for the resulting artificial dataset.

If a sample is already fused with the other two from the other modalities, it is not considered anymore for including no repetition of samples in the dataset. Data that could not be merged into trimodal examples are merged into bimodal or left unimodal.

The final artificial dataset contains only 141 points with the three modalities and 3426 that are incomplete (only one or two modalities). This is translated in few trimodal samples for base models (that requires complete data), and an acceptable number of samples for strategies dealing with missing data.

5.5. Fusion and recognition

For each approach, a “naive” strategy was implemented for comparing the same approach with and without dealing with missing data. The “naive” approaches consist of using the same approaches without dealing with missing data. Experiments were performed by approach. The metrics used for comparison are RMSE, $R^2$ score and std relative error. For every approach, RF, SVM and PLS regression models are compared, and for some, neural networks are implemented. Every model was tuned using a grid search approach.

5.5.1. First approach

Neural network models, RF, SVM and PLS models were experimented for this approach. The final neural network model used is a dense layer with a relu activation followed by a dropout layer and another dense layer at the end, with a tanh activation for prediction. Dropout and regularization strategies were used for avoiding overfitting. The optimizer used in this case is the Adam optimizer. The loss function used was the part after 1- in the $R^2$ formula. Contrary to what one might think, RF showed better results than the neural network model for this approach.

5.5.2. Second approach

For this approach, neural network models, RF, SVM and PLS models were explored. The final neural network model used is 3 LSTM layers with a relu activation each followed by a dropout layer. At the end, a dense layer with a tanh activation is used for prediction. Dropout and regularization strategies were used for avoiding overfitting. The optimizer used in this case is the Adam optimizer. The loss function used was mean squared error.

5.5.3. Third approach

As this approach is based on recursive neural network models, various architectures of this type of neural network were analyzed: simple recurrent, GRU and LSTM, obtaining the best results with the LSTM architecture. The final neural network model used is a LSTM layer with a relu activation followed by a dropout layer and a dense layer with a tanh activation for prediction. Dropout and regularization strategies were used for avoiding overfitting. The optimizer used in this case is the Adam optimizer. The loss function used was the part after 1- in the...
$R^2$ formula. Recurrent neural networks can receive varying length sequences as input, however, all sequences inside a batch must have the same length. For this reason, training and testing batches were formed appending randomly selected samples with equal length. All batches have a fixed size. 100 trimodal, 100 bimodal and 100 unimodal batches were formed for training and testing, random sampling is performed without replacement.

5.6. Models for predicting errors

The results of the approaches were not good enough. It was decided to add at the end of every approach a model for predicting errors of the fusion model, after that, predicted errors with predictions were summed for getting the final predictions, as proposed in [55]. Models considered for this method were SVM, RF and PLS. This method improved the performance of almost every approach, except for the second approach where no error model could improve the results of the fusion model, then, the fusion model is taken as the final model.

6. Results

Different results were obtained for the different unimodal models, and for the different multimodality fusion approaches dealing with missing data. Best results with unimodal models are for audio modality: $R^2 = 0.3$, RMSE = 0.36, SRE = 0.3; video: $R^2 = 0.07$, RMSE = 0.54, SRE = 0.39; and text: $R^2 = 0.43$, RMSE = 0.51, SRE = 0.25.

Audio modality performed better than the other two modalities, the reason for this can be found in the data or in the characteristics chosen to represent it. PLS regression was the best model for every one of the three modalities. SVM showed poor results in comparison with the other two techniques. RF showed comparable results to those of PLS in the test set, but they were biased to the center of the space, and low variance was present as evidenced by the SRE metric. In addition, huge overfitting was observed using this model. RF training performance was excellent, but in the testing set the results were very different, in contrast, PLS performed similarly in the training and the test sets.

"Naive" approaches (without dealing with missing data) gave different results to those of the proposed approaches. In fact, proposed approaches gave better results in general. Intuitively, "naive" approaches should perform better than proposed approaches because they do not have to deal with missing data. This lack of performance from "naive" approaches can be explained by the low quantity of complete data (only 141 samples) compared to the quantity of complete and incomplete data (3567 samples).

Table 3 resumes the results of "naive" and proposed approaches, and compares them with other works. Although the results of this work are not directly comparable with other works because we deal with missing data, the results from other works are exposed for having a reference.

As mentioned before, a comparison between our approaches and other works can not be directly performed because significant differences exist. Our approaches deal with missing data, also different datasets and different metrics are used. Works [27, 28] use Semaine as the dataset, we did not use this dataset because it does not contain text modality, thus, a speech-to-text process is required. Authors of [29] use AVEC 2017 dataset, this dataset is a subset of SEWA and contains the three modalities, where the text modality corresponds to the transcriptions of speech. This dataset was not used because it is in German, and is a bit different than the real conditions that happen in virtual learning environments. They obtain excellent results reaching an RMSE lower than 0.1, approximately. They use complex neural network models for extracting features from modalities and recognizing affective state, but those models are very time expensive and cannot be implementable in virtual education platforms due to scalability issues.

RF and PLS regression obtained the best results for almost every approach, and in some cases gave better results even than neural networks models. RF showed better results in terms of $R^2$ and RMSE for almost every approach, however, when predictions were plotted they were very central. Points far from the center of the space were never predicted, which was the reason for including SRE as a metric, for measuring the dispersion of predictions avoiding selecting models with high $R^2$ and RMSE but that are biased to the center of the space.

The base approach was included in the system to reference the results. However as mentioned before, they did not perform as expected. The performance of these base models is relatively close to the results of proposed approaches, noting differences that are relevant but not by much.

The best approach according to metrics is the first approach: feature-level fusion with zero padding using two RF models, one for predictions and one for predicting errors. Generally, RF showed centered biased results, and for this case, SRE is not so high showing a relatively good variance of predictions discarding center bias. In addition, $R^2$ is the highest observed.

These results showed that unimodal models do not contribute much to the system, as observed in the results of decision level fusion approaches, and that the direct fusion of the extracted characteristics is more successful than having an intermediate model. In fact, the unimodal models did not achieve encouraging results except for the audio modality.

7. Conclusions

In this work, an unobtrusive multimodal system for recognizing affective states in terms of arousal and valence that deals with missing
Table 3. Summary of comparison of continuous emotion approaches (B: base approach; F: First approach, feature level fusion with zero padding; S: Second approach, decision level fusion with zero padding; T: Third approach: decision level fusion with dynamic input, wavg: weighted average) for each modality (A: audio, V: video, T: Text) with different metrics (N/A: not apply, MLE: mean linear error).

| Modality | Data model | A | RMSE | SRE | MLE |
|----------|------------|---|------|-----|-----|
| B "naive" | A,V,T | Artificial | wavg,PLS | 0.370 | 0.334 | 0.404 | N/A |
| F "naive" | A,V,T | Artificial | PLS,PLS | 0.300 | 0.355 | 0.192 | N/A |
| S "naive" | A,V,T | Artificial | PLS,PLS | 0.400 | 0.326 | 0.401 | N/A |
| T "naive" | A,V,T | Artificial | NN,PLS | 0.346 | 0.340 | 0.400 | N/A |
| B | A,V,T | Artificial | wavg,RF | 0.394 | 0.357 | 0.538 | N/A |
| F | A,V,T | Artificial | RF,RF | 0.443 | 0.345 | 0.300 | N/A |
| S | A,V,T | Artificial | NN | 0.387 | 0.371 | 0.259 | N/A |
| T | A,V,T | Artificial | NN,NN | 0.281 | 0.409 | 0.474 | N/A |
| [27] | V | Semaine | SVM | N/A | 0.310 | N/A | N/A |
| [28] | AV | Semaine | SVM | N/A | N/A | N/A | 0.190 |
| [29] | A,V,T avec 2017 | NN | N/A | <0.1 | N/A | N/A |

Declarations

Author contribution statement

C. Salazar, E. Montoya-Múnera, J. Aguilar: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data will be made available on request.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] Derick Leon, et al., A widget to recommend learning resources based on the learner affective state, in: 3rd International Workshop on Motivational and Affective Aspects in Technology Enhanced, Learning (MA-T.E.L.2012), CEUR-Workshop Proceedings, 2012.

[2] Ana Raquel Faria, et al., A global perspective on an emotional learning model proposal, Telemat. Inform. 34 (6) (2017) 824837.

[3] Somayeh Fatahi, An experimental study on an adaptive e-learning environment based on learner’s personality and emotion, Educ. Inf. Technol. 24 (4) (2019) 2225–2241.

[4] Yi Yang, et al., Heart rate variability reflects the effects of emotional design principle on mental effort in multimedia learning, Comput. Hum. Behav. 89 (2018) 40–47.

[5] Reinhard Pekrun, The impact of emotions on learning and achievement: towards a theory of cognitive/motivational mediators, Appl. Psychol. 41 (4) (1992) 359–376.

[6] Liping Shen, Minjuan Wang, Ruimin Shen, Affective e-learning: using “emotional” data to improve learning in pervasive learning environment, J. Educ. Technol. Soc. 12 (2) (2009) 176–189.

[7] Liang-Chih Yu, et al., Improving early prediction of academic failure using sentiment analysis on self-evaluated comments, J. Comput. Assist. Learn. 34 (4) (2018) 358–365.

[8] Reinhard Pekrun, et al., Academic emotions in students’ self-regulated learning and achievement: a program of qualitative and quantitative research, Educ. Psychol. 97 (2) (2002) 91–105.

[9] Yuan-Pin Lin, et al., EEG-based emotion recognition in music listening, IEEE Trans. Biomed. Eng. 57 (7) (2010) 17981806.
Weifeng Liu, et al., Reinforcement online learning for emotion prediction using physiological signals, Pattern Recognit. Lett. 107 (2018) 123–130.

Panagiotis C. Perrantonakis, Leonios J. Hadjileontiadis, Emotion recognition from EEG using higher order crossings, IEEE Trans. Inf. Technol. Biomed. 14 (2) (2009) 186–197.

Morteza Zangeneh-Soroush, et al., A novel method of eeg-based emotion recognition using nonlinear features variability and Dempster-Shafer theory, Biomed. Eng.: Appl. Basis Commun. 30 (04) (2018) 185/026.

Jean Kossaif, et al., AFW-VA database for valence and arousal estimation in-the-wild, Image Vis. Comput. 65 (2017) 23–36.

Paul Ekman, Basic emotions, in: Handbook of Cognition and Emotion, vol. 98.45–48, 1999, p. 16.

Leila Kerkeni, et al., A review on speech emotion recognition: case of pedagogical interaction in classroom, in: 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATISP), IEEE, 2017, pp. 1–7.

Keith Nesbitt, et al., Using the stare eye-blink to measure affect in players, in: Serious Games Analytics, Springer, 2015, pp. 401–434.

Soujanya Poria, et al., A review of affective computing: from unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98125.

Camilo Salazar, et al., Affective recommender systems in the educational field: A systematic literature review, Comput. Sci. Rev. 40 (2021) 100377.

Sergio Salmeron-Majadas, et al., Filtering of spontaneous and low intensity emotions in educational contexts, in: International Conference on Artificial Intelligence in Education, Springer, 2015, pp. 429–438.

Sergio Salmeron-Majadas, Olga C. Santos, Jesus G. Boticario, An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context, Proc. Comput. Sci. 35 (2014) 691–700.

Margaret M. Bradley, Peter J. Lang, Measuring emotion: the self-assessment manikin and the semantic differential, J. Behav. Ther. Exp. Psychiatry 25 (1) (1994) 49–59.

Keng-hao Chang, Drew Fisher, John Canny, Ammon: a speech analysis library for analyzing affect, stress, and mental health on mobile phones, in: PhoneSense 2011, 2011.

Bjorn Schuller, et al., The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, in: 14th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2013, pp. 148–152.

Soujanya Poria, et al., Towards an intelligent framework for multimodal affective data analysis, Neural Netw. 63 (2015) 104–116.

Soujanya Poria, et al., Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: IEEE 16th International Conference on Data Mining (ICDM), IEEE, 2016, pp. 439–448.

Carlos Busso, et al., IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335.

Hisato Gunes, Maja Pantic, Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners, in: International Conference on Intelligent Virtual Agents, Springer, 2010, pp. 371–377.

Florian Eyben, et al., String-based audiovisual fusion of behavioural events for the assessment of dimensional affect, in: Face and Gesture, 2011, pp. 322–329.

Shizhe Chen, et al., Multimodal multi-task learning for dimensional and continuous emotion recognition, in: 7th Annual Workshop on Audio-Visual Emotion Challenge, 2017, pp. 19–26.

Johannes Wagner, et al., Exploring fusion methods for multimodal emotion recognition with missing data, IEEE Trans. Affect. Comput. 2 (4) (2011) 206–218.

G. Candikas, et al., A multimodal corpus for gesture expressivity analysis multimodal corpora, in: Workshop Multimodal Corpora, 2010, p. 80.

Yongqiang Li, et al., Handling missing labels and class imbalance challenges simultaneously for facial action unit recognition, Multimed. Tools Appl. 78 (14) (2019) 20309–20332.

Mara Araujo, Jose Aguilar, Hector Aponte, Fault detection system in gas lift well based on artificial immune system, in: International Joint Conference on Neural Networks, 2003, vol. 3, 2003, pp. 1673–1677.

Marvin Jimenez, et al., An automatic approach of audio feature engineering for the extraction, analysis and selection of descriptors, Int. J. Multimed. Inf. Retr. 10 (2021) 33–42.

Fanny Pacheco, et al., Methodological framework for data processing based on the Data Science paradigm, in: 2014 XL Latin American Computing Conference, 2014 (CLEI).

Davis E. King, Diib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.

Erik Cambria, et al., SenticNet 5: discovering conceptual primitives for sentiment analysis by means of context embeddings, in: Thirty-Second AAAI Conference on Artificial Intelligence, vol. 32, 2018, p. 17951802.

Erik Cambria, et al., Affectivespace 2: enabling affective intuition for concept-level sentiment analysis, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, vol. 29, 2015, pp. 508–514.

Mohammad Soleymani, et al., A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (1) (2011) 42–55.

Soujanya Poria, et al., EmsSenticSpace: a novel framework for affective common-sense reasoning, Knowl.-Based Syst. 69 (2014) 108123.

Daniel Preotiuc-Pietro, et al., Modelling valence and arousal in Facebook posts, in: Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2016, pp. 9–15.

Olivier Martin, et al., The eNTERFACE’05 audio-visual emotion database, in: 22nd International Conference on Data Engineering Workshops (ICDEW’06), IEEE, 2006, 8 pp.

Amir Zadeh, Ali Bagher, et al., Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.

Gary McKeown, et al., The Semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent, IEEE Trans. Affect. Comput. 3 (1) (2011) 517.

Amir Zadeh, et al., Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88.

Veronica Perez Rosas, Rada Mihalcea, Louis-Philippe Morency, Multimodal sentiment analysis of Spanish online videos, IEEE Intell. Syst. 28 (3) (2013) 38–45.

Soujanya Poria, et al., MILD: a multimodal multi-party dataset for emotion recognition in conversations, arXiv preprint, arXiv:1810.02508, 2018.

Ellen Douglas-Cowie, et al., The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data, in: International Conference on Affective Computing and Intelligent Interaction, Springer, 2007, pp. 488–500.

N. Perez, et al., A Verification Method for MASORS, IEEE Trans. Cybern. 43 (1) (2013) 64–76.

Takeso Kanade, Jeffrey F. Cohn, Yingli Tian, Comprehensive database for facial expression analysis, in: Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), IEEE, 2000, pp. 46–53.

Patrick Lucey, et al., The Extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specification expression, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, IEEE, 2010, pp. 94–101.

Sander Koolstra, et al., Deap: a database for emotion analysis; using physiological signals, IEEE Trans. Affect. Comput. 3 (1) (2011) 18–31.

Juan Correa, Miranda Abdon, et al., AMIGOS: a dataset for affect, personality and mood research on individuals and groups, IEEE Trans. Affect. Comput. (2018).

E. Puerto, et al., Using multilayer fuzzy cognitive maps to diagnose autism spectrum disorder, Appl. Soft Comput. (ISSN 1568-4946) 75 (2019) 5871.

Sushil Punia, et al., Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail, Int. J. Prod. Res. 58 (16) (2020) 49644979.