A Multi-modal Visual Emotion Recognition Method to Instantiate an Ontology

Juan Heredia, Yudith Cardinale, Irvin Dongo, Jose Díaz-Amado

A Multi-modal Visual Emotion Recognition Method to Instantiate an Ontology

Juan Pablo A. Heredia[1][a], Yudith Cardinale[1,2][b], Irvin Dongo[1,3][c] and Jose Díaz-Amado[1,4][d]
[1]*Computer Science Department, Universidad Católica San Pablo, 04001 Arequipa, Peru*
[2]*Computer Science Department, Universidad Simón Bolívar, 1080 Caracas, Venezuela*
[3]*Univ. Bordeaux, ESTIA Institute of Technology, 64210 Bidart, France*
[4]*Electrical Engineering, Instituto Federal da Bahia, 45078-300 Vitoria da Conquista, Brazil*

Keywords: Emotion Recognition, Multi-modal Method, Emotion Ontology, Visual Expressions.

Abstract: Human emotion recognition from visual expressions is an important research area in computer vision and machine learning owing to its significant scientific and commercial potential. Since visual expressions can be captured from different modalities (e.g., face expressions, body posture, hands pose), multi-modal methods are becoming popular for analyzing human reactions. In contexts in which human emotion detection is performed to associate emotions to certain events or objects to support decision making or for further analysis, it is useful to keep this information in semantic repositories, which offers a wide range of possibilities for implementing smart applications. We propose a multi-modal method for human emotion recognition and an ontology-based approach to store the classification results in EMONTO, an extensible ontology to model emotions. The multi-modal method analyzes facial expressions, body gestures, and features from the body and the environment to determine an emotional state; this processes each modality with a specialized deep learning model and applying a fusion method. Our fusion method, called EmbraceNet+, consists of a branched architecture that integrates the EmbraceNet fusion method with other ones. We experimentally evaluate our multi-modal method on an adaptation of the EMOTIC dataset. Results show that our method outperforms the single-modal methods.

1 INTRODUCTION

In human communication, non-verbal messages contain lots of information, including emotional state, attitude, or intentions of the people (Knapp et al., 2013). Thus, human emotion recognition from visual expressions is an important research area of computer vision and deep learning owing to its significant academic, scientific, and commercial potential (Zhang et al., 2018).

Most of the research related to visual expressions analysis for emotion recognition refers to facial gestures, leaving those of the rest of the body a little aside (Lhommet et al., 2015). However, the analysis of other visual factors, such as body posture, the move of limbs, clothes or breath, or the context, represents an alternative or complement to improve emotion recognition results (Noroozi et al., 2018). In particular, analyzing the body posture is important, because it can express emotional states involuntarily and spontaneously (Knapp et al., 2013; Cowen and Keltner, 2020). Similarly, the physical environment or context matters because it can influence or condition the people emotional state (Knapp et al., 2013; Mittal et al., 2020). This context can communicate about human-object and human-human interactions that occur indirectly or separately (Cowen and Keltner, 2020).

In this sense, since people naturally express emotions in simultaneous different ways (i.e., face expressions, hands posture, body posture) that can be conditioned by their context, multi-modal methods are suitable and are becoming popular for emotion recognition (Noroozi et al., 2018; Kaur and Kautish, 2019). Multi-modal methods, generally based on deep learning models, simultaneously analyze the different visual expression and factor modalities to identify hu-

[a] https://orcid.org/0000-0002-6126-4881
[b] https://orcid.org/0000-0002-5966-0113
[c] https://orcid.org/0000-0003-4859-0428
[d] https://orcid.org/0000-0001-8447-784X

Heredia, J., Cardinale, Y., Dongo, I. and Díaz-Amado, J.
A Multi-modal Visual Emotion Recognition Method to Instantiate an Ontology.
DOI: 10.5220/0010516104530464
In *Proceedings of the 16th International Conference on Software Technologies (ICSOFT 2021)*, pages 453-464
ISBN: 978-989-758-523-4

man reactions in different research contexts (e.g., human-computer interaction, healthcare, extracting opinions about events or objects) (Kaur and Kautish, 2019; Egger et al., 2020).

These multi-modal methods can overcome the limitations of performing emotion recognition using only one kind of signal or gesture, getting more robust results. Nevertheless, they have other limitations for recognizing emotions, related to the dataset used, the modality selection, the computational requirements for a proper execution, and the modalities synthesis or fusion. The data is a limitation because deep learning methods need a lot of data for learning. Moreover, most datasets are focused and designed for only one modality, commonly facial expressions (Kaur and Kautish, 2019). Therefore, a data pre-processing on the original samples should be used to obtain the data of various types (i.e., the modality data); even so, the presence of all modalities is not guaranteed, which is a problem that should be solved in the training phase. The selection of modalities depends on the purpose of the application and the data available, and typically, the model demands having all selected modalities to perform ideally. The fusion of modalities is a very important aspect and must ensure that all available data are used to the maximum (Soleymania et al., 2017). The methods for merging the modalities are based on probabilities (e.g., EmbraceNet (Choi and Lee, 2019a)), simple or complex trainable decisions (e.g. Multiplicative Fusion (Liu et al., 2018)), another learning phase (e.g., Multiple Kernel Learning), and so on. Most of these methods can be implemented with any deep learning multi-modal method, but the same quality of results is not assured for all (Dong et al., 2020; Egger et al., 2020).

To overcome these limitations, we propose a multi-modal method able to analyze facial expressions, body gestures, and features from the body and the environment from images, to estimate the emotional state of a person. Nevertheless, it is not mandatory to analyze all modalities; the model handles the absence of some of the modalities (e.g., hidden face). To recognize an emotion, each type of cues is processed in a specialized and independent deep learning model. Then, a fusion method is applied, which consists of a branched architecture that integrates multiple fusion methods. This fusion method, called EmbraceNet+, extends the EmbraceNet fusion method by altering its structure and integrating it into an architecture that allows three EmbraceNets to be used with any other fusion method.

Since the analysis of human emotions is generally performed to associate emotions to certain events or objects and consequently to make decisions or for further analysis, keeping this information in semantic repositories offers a wide range of possibilities for implementing smart applications (Bertola and Patti, 2016; Chen et al., 2016; Cavaliere and Senatore, 2019; Graterol et al., 2021).

To support decision making and post-analysis from the analysis of emotions, we use an emotion ontology, called EMONTO (Graterol et al., 2021). EMONTO is an extensible emotion ontology, that can be integrated with other specific domain ontologies. Therefore, it is possible to combine the semantic information of emotions with other ontologies, that represent entities to which the emotions can be associated – e.g., emotions produced by artworks in museums or by the food in restaurants or by candidates in government elections.

We experimentally evaluate our multi-modal method in an adaptation of the EMOTIC dataset (Kosti et al., 2017; Kosti et al., 2019); showing that the combination and integration of various fusion methods allow beating the results from single-modal methods and even beat those obtained from using only EmbraceNet as fusion method. Furthermore, the results obtained are competitive and similar to those reported from the state-of-the-art methods: EMOTIC model (Kosti et al., 2019) and EmotiCon (Mittal et al., 2020).

In summary, the contribution of this work is threefold: (i) a multi-modal method to recognize human emotions from images in any situation, which adapts to the present modalities (i.e., face expression, body posture, features from body and context); (ii) EmbraceNet+, a fusion method that, like EmbraceNet, can be used in any multi-modal machine learning-based model; and (iii) the integration of an extensible emotion ontology (EMONTO) and the approach to instantiate it; this semantic repository can be combined with other domain ontologies to relate emotions in any context.

2 RELATED WORK

In this section, we survey some studies on visual emotion recognition using deep learning, common approaches, and recent achievements on fusion models; as well as the studies on emotion recognition based on ontologies.

2.1 Visual Emotion Recognition

The common visual emotion recognition from human faces began as an exploration of deep learning models, such as Convolutional Neural Networks (CNN),

focused on the face image region, as in (Parkhi et al., 2015). However, now some additional aspects are considered to improve the results. For instance, (Zadeh et al., 2019) use a Gabon filter at the beginning as an extra feature extractor. Recent deep learning models, such as Graph CNN (GCNN), have gained popularity in the analysis of body movements and postures to detect human emotions.

Moreover, the context analysis started as a recognition of the emotions that an image expresses by itself. Currently, the context is used as a complement of people data for achieving competitive results, even in situations where the people data is unclear or inappropriate. For example, the approach proposed by (Lee et al., 2019) uses the whole image without the facial region, by processing the context with an attention-perception model. This new type of approach is already considered a multi-modal one.

Most multi-modal methods differ mainly in the modalities of input data used, and in the method to fusion the modalities. The method proposed by (Kosti et al., 2019) uses two kinds of data: the whole image and a sub-image of the person. This model processes the inputs with two CNN models, whose results are concatenated and processed again with a Multi-Layer Perceptron (MLP) neural network. (Mittal et al., 2020) propose a work that uses up to four different modalities: the face landmarks, a graph of the human posture, the whole image without the region delimited by the person, and a depth mask. The human posture graph is processed by a GCNN model and the rest of modalities by different CNN models. The fusion of this model consists on merging the results of facial and postural modalities with the Multiplicative Fusion technique, then, similar to the work proposed by (Kosti et al., 2019), it concatenates the remained three results and use an MLP neural network to get the final prediction. As well as the concatenation, the Weighted Sum (WS) is commonly used because it is very simple to apply and does not need many resources (Dong et al., 2020).

The approaches about multi-modal emotion recognition have proven be effective applications in real-world data (Zhang et al., 2019). For instance, in robotic, the studies presented by (Perez-Gaspar et al., 2016; Chumkamon et al., 2016; Graterol et al., 2021) are based on a multimodal emotion recognition, understanding the reality with data coming from the robots' sensors, to then make decisions from the data processed. Other intelligent systems that use the emotion analysis, could be benefited with a multi-modal method, such as the music recommender proposed by (Gilda et al., 2017), which selects songs according to the listener mood, or the abnormal human behavior recognizer presented by (Caruccio et al., 2019), which uses several contexts, including one related to the recognition of emotions, to improve the accuracy of the results.

2.2 Emotion Recognition and Ontologies

Some studies combine emotion recognition with ontologies. However, most of them use the ontologies to support the classification process on recognizing emotions. Besides that, they are limited, either to a specific modality (e.g., text, face, body) or a specific source (e.g., text, images, video). For example, the approaches to recognize emotions from texts in social media proposed by (Chen et al., 2016; Cavaliere and Senatore, 2019), are based on their own emotion ontologies to support the classification process. These works are not suitable to establish relations between emotions and objects or events, since they are not able to represent such entities. An approach to perform sentiment analysis on social media to relate sentiment to artworks is proposed by (Bertola and Patti, 2016). However, the applicability of this approach cannot be extended to other domains, as our proposal.

Regarding emotions modeling, there are many ontologies aimed at representing emotions. However, most of them are domain dependent or emotion model dependent, such as the work presented by (Zhang et al., 2013), whose ontology is dedicated to represent electroencephalographic (EEG) data and applied it to detect human emotional states; the framework presented by (Garay-Vitoria et al., 2019), based on an emotion ontology, applicable for developing affective interaction systems; or the ontology proposed by (Lin et al., 2018b) to represent emotions with visual elements. Other emotion/sentiment ontologies model an emotion as a main entity which has associated modalities, dimensions, categories, and other features (Grassi, 2009; Lin et al., 2018a), and also fuzzy concepts (Ali et al., 2017; Dragoni et al., 2018), but they are still restricted and limited in applicability and extensibility, due to the lack of connection with entities that can connect values from other domains (e.g., an object entity can be instantiated as an artwork in the tourism domain). Based on the previous limitations, a recent general ontology, called EMONTO (Graterol et al., 2021), has been proposed to represent several sources, modalities, different emotion models, and also be extended with specific domain ontologies to represent objects/events to which sentiment can be related.

Since our aim is to support the development of further applications able to analyze semantic informa-

tion, independently of the domain, the modalities, and the emotion classification, we adopt EMONTO as the model to store the recognized emotions into a semantic repository.

3 MULTI-MODAL EMOTION RECOGNITION

The whole pipeline of our proposal consists of four phases, as shown in Figure 1. The first three are about the multi-modal emotion recognition method: (i) the extraction of data for each modality; (ii) emotion recognition using independent methods that deal with each modality; and (iii) the fusion of modalities to obtain a multi-label classification. The last one, is the phase of instantiation of the semantic repository, which is developed in Section 4.

3.1 Proper Data Acquisition

The input of the proposed method are images, which ideally should have at least one person. The process for obtaining the appropriate data (i.e., body image, face image, context image, and joint and limb graph) is as follows:

1. Each person within the image is recognized using a computer vision model like YOLO (Farhadi and Redmon, 2018), which provides a list of bounding boxes with each person's location.
2. Using the bounding boxes, sub-images of people are extracted. Moreover, the **context image** is getting by removing everything in the bounding boxes. For each bounding box, a sub-image and a context image is generated.
3. The sub-images of persons are processed using two deep learning models. One of them is the HR-Net (Sun et al., 2019) to human pose estimation that gives the location of every visible person's joint; with these data, the **joint and limb graphs** that encode the body gestures are generated. The other model is the RetinaFace (Deng et al., 2019) that provides several facial information, however only the face bounding box is used.
4. Another sub-image is extracted using the face bounding box, and as well as with the whole image, the face is removed from the person image, which conform the **body image**. The face sub-image is the **face image**.

The reshaping of the four vectors obtained is implementation-free, but it must be specific shapes: the body and context images ($224 \times 224 \times 3$); the face image (48×48); and the joint and limb graphs ($Nchanels \times Nframes \times Njoints$ or $Nlimbs$), where $Nchanels$ is the x and y axises position and the point confidence, $Nframes$ is the number of frames set to 1, and $Njoints$ and $Nlimbsset$ are the numbers of joints and limbs, both set to 15 due the HRnet model pre-trained in MS-COCO.

3.2 Independent Processing

The proposed multi-model method processes each modality in particular ways:

Facial Modality. It is processed with an adaptation of the VGG-face model (Parkhi et al., 2015). The same architecture is used, but the final MLP is modified, since the dimensions of the input images, set to 48×48, cause the number of end features to decrease to 512. Therefore, the fully connected layers become just one of 512-N neurons – i.e., there are 512 input neurons and N output neurons, N is the number of emotions. The new VGG-face consists of five convolutional blocks and the final MLP. The convolutional blocks have two convolutional layers of 64 filters, two of 128, four of 256, and four of 512 filters. At the end of each convolutional block, there is a max pooling layer. Each convolutional layer also contains a batch

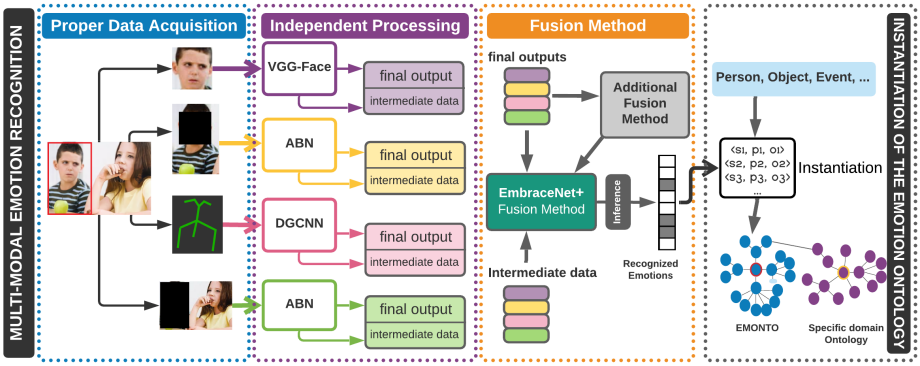

Figure 1: Structure of the entire proposal.

normalization and a ReLU activation function.

Postural Modality. Joint and limb graphs represent the human posture that capture body gestures and allows analyzing the movement and dynamics of the body's limbs. We use the Directed GCNN (DGCNN) model (Shi et al., 2019) to process this modality. Direct graphs allow the model to learn about the dependency and dynamics of joints and limbs. The DGCNN model is based on graph-temporal convolutions, which contain DGCNN blocks, bi-temporal convolutions, a ReLU layer, and a residual layer that in some cases are bi-temporal convolutions. The bi-temporal convolution consists of two convolutions on the temporal dimension. The architecture of this model is composed of ten layers of graph-temporal convolution, the first four of 64 filters, the next three of 128, and the last three of 256 filters; at the end, an MLP that outputs the final classification (Shi et al., 2019).

Contextual and Body Modalities. For both context and body modalities the same model is used for their processing: the Attention Branch Network (ABN) (Fukui et al., 2019). It is based on an attention mechanism to improve the visual explanation and image recognition with CNN. The ABN is made up of three components: a feature extractor, the attention branch, and the perception branch; the attention mechanism works as a linker of the three main components.

For our proposed method, the idea of removing the persons from the image to process the context, presented by (Mittal et al., 2020) in EmotiCon, is followed. However, they delete everything within the person bounding box, losing body features such as clothing, held objects, or other items. In contrast, we process the body as another modality. The body is processed in the same way the context is processed but removing the face in this case. Thus, our model can recognize relevant objects or agents that can influence the emotional state.

3.3 Fusion Method: EmbraceNet+

The fusion method is usually one of the main contributions of multi-modal approaches. In our proposed fusion method, the EmbraceNet model (Choi and Lee, 2019a) is taken as the basis. EmbraceNet can be applied to merge intermediate data and then apply an additional model, or to merge the final results and then, if it is necessary, apply an additional model. Moreover, thanks to the versatility of EmbraceNet, there are other ways to implement it, being able to use more than one EmbraceNet or even with other fusion methods – e.g., weighted sums or votes that already process characteristics and can provide additional relevant information (Dong et al., 2020).

The proposed fusion method, called EmbraceNet+, is a branched architecture that allows using more than one EmbraceNet together with other fusion methods (see Figure 2). It has two EmbraceNets that deal with intermediate data and final outputs from every modality, respectively; and then, their results are the input for a third EmbraceNet, along with other fusion methods. The first two EmbraceNets, and each of the additional fusion methods used, can be viewed as branches.

Every EmbraceNet is composed by docking layers and an embrace layer (Choi and Lee, 2019a). The docking layers have two functions: transform the inputs from each modality into a same shape and learn about the correlation among modalities, they learn from the features of every modality. Sometimes, the input vectors already have the same shape, or they have a few features; thus, an MLP of one layer could be not enough to get a significant improvement re-

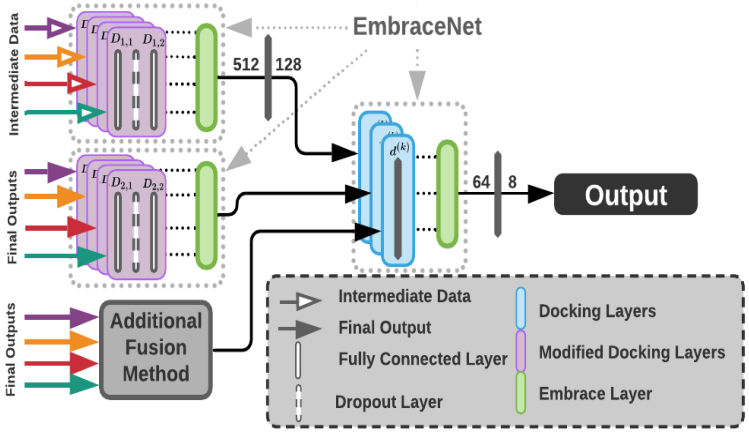

Figure 2: EmbraceNet+ Architecture.

garding single modality methods. For this reason, we also propose a modification over the docking layers of the two EmbraceNets that directly process the modalities results. This modification consists in adding two layers to the MLP in docking layer: a fully connected and a dropout; it is expected that the docking layers improve their modalities correlation and work as additional feature extractors to improve the final result. Thus, the new docking MLP are composed by:

1. A fully connected layer of $I^{(k)}$-$D^{(k)}_{x,1}$, where (k) indicates the modality coming from the k^{th} modality and x the container EmbraceNet, $x = 1$ if it deals with intermediate data and $x = 2$ if it is the second one. I represents the intermediate data feature number, in our case everyone is 512 or N(number of emotions), therefore, all $D_{*,1}$ are also equal and they are set as $D_{1,1} = 1024$ and $D_{2,1} = 64$.
2. A dropout layer, mainly to avoid an overfiting in the training phase. This layer is set with 0.5 as the factor of drop. The addition of this layer also responds to the proposed improvements to the original EmbraceNet, where a training with random dropout of modalities is present for strengthen the method.
3. A fully connected layer of $D^{(k)}_{x,1}$-$D_{x,2}$ that mainly must set each vector in a same shape; then, for the first EmbraceNet is set $D_{1,2} = 512$ and $D_{2,2} = 32$. The values of each docking layer were set by experimentation.

The proposed method also has as strength the using of another fusion strategies. We propose applying over the modalities results a WS and a concatenation, then both are used together with the prior EmbraceNets output, as the input to the third EmbraceNet. Before the first EmbraceNet output go to the final merge, it is applied an extra MLP of 512-128, to avoid affecting the training of the final fusion because the dramatic difference in the number of features. Likewise, the docking layers of the final EmbraceNet are not modified; they are four and their configurations are 128-64, 32-64, 32-64, and 8-64, to deal with the first Embracenet, the second one, the concatenation and the WS, respectively.

Respecting the embrace layers, the original is used for every EmbraceNet without modification. This layer is based on the multinomial sampling, and it is defined by the Equation (1), where e_i is the embraced result for the i^{th} feature; the $r^{(k)}$ vector of k^{th} modality is obtained by $r^{(k)}_i \sim Multinomial(1, p)$, where p is a defined probability, there are as many p as modalities, such that $\sum_k p^{(k)} = 1$; $d^{(k)}_i$ represents the result of the docking layers. This process ensures that only one modality contributes to i^{th} component of the embraced vector. Those probabilities from EmbraceNets, and of the WS, are set the same for all, $1/n_{modalities}$, other distributions were probed but a significant improvement was not found.

$$e_i = \sum_k r^{(k)}_i \cdot d^{(k)}_i \tag{1}$$

The proposed method deals with missing or wrong data by using availability values: 1 if there are data for the modality, 0 otherwise. There are as many values as modalities. The absence of data is detected in the initial data passing; for example, for posture modality, a threshold of confidence is set to determinate if the posture estimation is favorable. When there is a 0 in availability values, internal probabilities, of each EmbraceNet and WS, are adjusted, maintaining the same distribution but without considering the missing modality. For obtaining the final inference, the output represented as a real number vector is converted to a binary vector, based on thresholds. These thresholds could be defined in the validation or testing phases, based on the precision-recall curve. The implementation of the multi-modal model is available in the GitHub repository https://github.com/juan1t0/multimodalDLforER.

4 EMOTION ONTOLOGY AND INSTANTIATION

The data provided by emotion recognition methods, such as emotion labels, its probabilities, modalities,

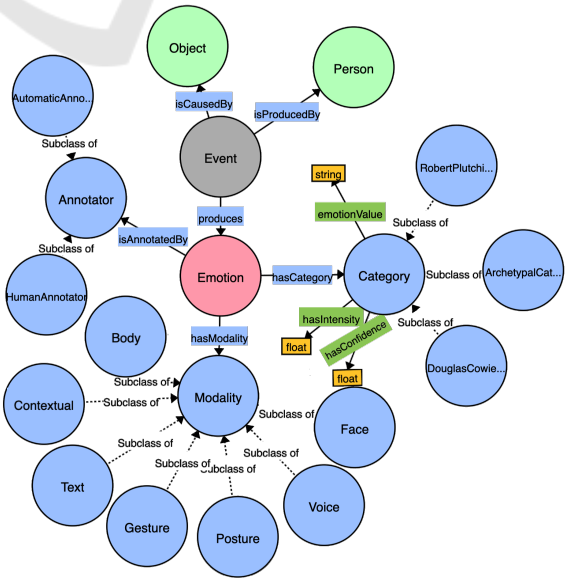

Figure 3: EMONTO: An Extensible Emotion Ontology.

source, as well as the entities to which the emotions are related, such as Person, Object, Place, are normally stored for further analysis and decision making, which can generate complex networks of knowledge. In this sense, it is evident the necessity of a well-defined and standard model, such as ontologies, for representing the huge quantity of knowledge managed by applications that produce real-time values. In the following, we describe EMONTO (Graterol et al., 2021), the ontology that is integrated at the end of the pipeline of our multi-modal emotion recognition process.

4.1 EMONTO: An Extensible Emotion Ontology

EMONTO is an extensible ontology that represents emotions under different categorization proposals. It adopts some concepts from several emotion ontologies (Grassi, 2009; Lin et al., 2018a) to provide compatibility. Figure 3 shows the main classes of EMONTO. The central class is `emo:Emotion`, which has a category (`hasCategory`) according to a *Category* class. Currently, EMONTO considers Archetypal (Grassi, 2009), Douglas Cowie (Grassi, 2009), and Robert Plutchik (Plutchik, 1980) emotion categorizations, which group emotions into 6, 25, and 8 values, respectively. Nevertheless, any other categorization model can be integrated as a subclass of `emo:Category`.

EMONTO has `emo:Event` as a class that connects `emo:Object`, `emo:Person`, and `emo:Emotion` entities. An emotional *Event* is produced by (`isProducedBy`) a *Person* and also it is caused by (`isCausedBy`) an *Object* (e.g., artworks, candidates, plates). An *Event* can produce several *Emotions*.

The entities `emo:Object` and `emo:Person` are general classes that can connect other ontologies, such as museum and artwork ontologies (Pinto-De la Gala et al., 2021) as *Object* or user-profile ontologies (Katifori et al., 2007) as *Person*. EMONTO is extensible and flexible to be easily adopted in scenarios where data related to the recognized emotions, need to be stored for further analysis. The ontology provides the modality (`emo:Modality`) of the information used to recognize the emotion (e.g., `emo:Gesture`, `emo:Face`, `emo:Posture`), and the type of annotator (`emo:AutomaticAnnotator` and `emo:HumanAnnotaton`). Moreover, datatype properties `emo:hasIntensity` and `emo:hasConfidence` are associated to the category to express the level of intensity and confidence (probability score), respectively; float values between 0 and 1.0.

4.2 Instantiation of the Emotion Ontology

Once the emotions are recognized in the previous phases of the pipeline (see Figure 1), an *Event* is created to represent the emotional event, which consists of a set of emotions produced at a certain time. For example, let's consider *anger* and *disgust* as the detected emotions (`EM0001` and `EM0002`, respectively) from the *Douglas Cowie* category; therefore, an emotional event is created (`<EV0001 rdf:type emo:Event>`) and associated to some datatype properties – e.g., `<EV0001 emo:createdAt 1612748699>`, where the object value is a Unix Timestamp (date). Afterward, the *Event* is associated to a *Person* and *Object* (e.g., `<EV0001 emo:isProducedBy P0001>` and `<0001 emo:isCausedBy O0001>`, where `P0001` represents a *Person* and `O0001` represents the *Object* that caused the *Emotion* event `EV0001`).

Person and *Object* should be also recognized by using other models that can be combined with our multi-modal method in the **proper data acquisition** phase. For example, by applying face recognition to identify registered or new users in the system and object detection to recognize specific objects (e.g., artworks in museums, plates in restaurants). This work is focused on the emotion recognition, *Person* and *Object* detection is beyond the scope of this research.

According to the results of the emotion recognition method, which are *anger* and *disgust* in the previous example, the emotion entities are created (`<EM0001 rdf:type emo:Emotion>` and `<EM0002 rdf:type emo:Emotion>`) and associated to the *Event* (`<EV0001 emo:produces EM0001>` and `<EV0001 emo:produces EM0002>`). Each emotion has a *Category* (`<EM0001 emo:hasCategory C0001>` and `<EM0002 emo:hasCategory C0002>`) which can be Archetypal, Douglas Cowie, or Robert Plutchik classifications (`<C0001 rdf:type ArchetypalCategory>`, `<C0002 rdf:type ArchetypalCategory>`). The datatype properties `emo:hasConfidence` and `emo:emotionValue` are added to the *Category* (`<C0001 emo:hasConfidence 0.13>`, `<C0001 emo:emotionValue "anger">` and `<C0002 emo:hasConfidence 0.35>`, `<C0002 emo:emotionValue "disgust">`) with the values obtained by the emotion recognition method (e.g., *0.13* and *0.35* as confidence - probabilities; *"anger"* and *"disgust"* as emotions). *Modality* and *Annotator* are also instantiated.

Algorithm 1 presents the general process of the ontology instantiation. First, libraries related to the RDF management have to be import (line 1), then a new Graph, which contains the RDF triples is cre-

ated (line 2). Namespaces of the ontology are added (lines 3-5). From line 6 to line 15, new RDF triples are added to the Graph.

Algorithm 1: Creating RDF triples.

```
1  import RDF libraries
2  g = Graph()
3  EMO = Namespace("http://www.emonto.org/")
     //Creating a Namespace.
4  g.add_namespace("emo", EMO)  //Adding the
     namespace EMO.
5  g.add_namespace("foaf", FOAF)  //Adding the
     namespace FOAF, which is already defined in the
     libraries.
6  g.add_triple((EMO.EV0001, RDF.type,
     EMO.Event))  //Creating an Event.
7  g.add_triple((EMO.P0001, RDF.type,
     FOAF.Person))  //Creating a Person.
8  g.add_triple((EMO.O0001, RDF.type,
     EMO.Object))  //Creating a Object.
9  g.add_triple((EMO.EM0001, RDF.type,
     EMO.Emotion))  //Creating an Emotion 1.
10 g.add_triple((EMO.EM0002, RDF.type,
     EMO.Emotion))  //Creating an Emotion 2.
11 g.add_triple((EMO.EV0001, EMO.produces,
     EMO.EM0001))  //Associating EV0001 to
     EM001.
12 g.add_triple((EMO.EV0001, EMO.produces,
     EMO.EM0002))  //Associating EV0001 to
     EM002.
13 ...
14 g.add_triple((EMO.C0001, EMO.hasConfidence,
     Literal(0.13)))  //Confidence value.
15 g.add_triple((EMO.C0001, EMO.emotionValue,
     Literal("anger")))  //Recognized emotion.
16 ...
```

Algorithm 2: Obtaining emotions of artwork Mona_Lisa.

```
1  import RDF libraries
2  g = Graph()
3  g.read("database_emotions.ttl", format="ttl")
4  g.add_namespace("emo",
     "http://www.emonto.org/")
5  qres = g.query( "SELECT ?emotion_l
6                WHERE {
7                ?emotion a emo:Emotion ;
8                ?emotion emo:emotionValue
9                ?emotion_l .
10               ?event a emo:Event ;
11               ?event emo:produces ?emotion ;
12               ?event emo:isCausedBy
13               ?object.
14               ?object a emo:Object ;
15               ?object rdfs:label "Mona_Lisa" .
16 }")
```

The semantic repository can be queried later for specific information, such as artworks and places in the tourism domain. For instance, Algorithm 2 retrieves the emotions produced by the *Mona_Lisa* artwork. Libraries related to the RDF manipulation are imported (line 1); then, a Graph should be initialized to read the RDF triples (line 2 and line 3); the namespaces used in the ontology have to be added (line 4); then, following the SPARQL syntax, a query is performed (lines 5-15).

5 EVALUATION OF THE MULTI-MODAL METHOD

In this section, we show the performance of our multi-modal method compared with the EMOTIC model (Kosti et al., 2019), which was proposed along with the dataset EMOTIC, and EmotiCon (Mittal et al., 2020). We detail the procedure carried out to adequate the dataset used and the achieved results.

5.1 Dataset Adequacy

The EMOTIC database provides the emotion annotations along with the bounding box of the person who feels the emotion annotated. Therefore, the process of person recognition and location is omitted.

EMOTIC has 26 labeled emotions as discrete categories and almost no images have a single category; that is, the classification is a multi-class and multi-label problem. Some of those discrete categories are specific and not common; therefore, they may not provide relevant information in most real-world applications. Because of this, we relabel the dataset by grouping similar emotions into Category Groups (CG), using the taxonomy proposed in (Plutchik, 1980) and in accordance with the definition of each original emotion (Kosti et al., 2017). Those CG are: *Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise*, and *Trust*; the original emotions included in each CG are detailed in Table 1. Besides, the discrete categories are unbalanced; thus, a weighted random sampler was used in the training phase, oversampling some samples of the categories with less presence. In addition, a random crop function is applied to perform a data augmentation.

Table 1: Grouping of the original EMOTIC emotions into eight categories.

Cat. Group	Original Emotions
Anger	Anger, Annoyance, Disapproval
Anticipation	Anticipation, Engagement
Disgust	Aversion, Disconnection, Fatigue, Yearning
Fear	Disquietment, Embarrassment, Fear
Joy	Affection, Excitement, Happiness, Peace, Pleasure
Sadness	Pain, Sadness, Sensitivity, Suffering
Surprise	Doubt/Confusion, Surprise
Trust	Confidence, Esteem, Sympathy

Table 2: Results of the AP metric obtained by single-modal methods.

Cat. Group	Modalities			
	Body	Contextual	Facial	Postural
Anger	13.60	14.11	15.06	1.02
Anticip.	88.46	87.73	85.75	1.68
Disgust	30.34	31.89	24.63	3.25
Fear	17.29	17.06	15.86	9.21
Joy	81.28	81.47	79.80	82.22
Sadness	14.85	14.53	10.82	2.39
Surprise	23.04	22.40	19.99	4.10
Trust	71.43	69.29	57.82	10.50
mAP	42.54	42.31	38.72	14.30

5.2 Results

Our experiments are split in two: (i) comparisons with the *Average Precision* (AP) and *mean AP* (mAP) metrics with all **single-modal** and **multi-modal** methods; and (ii) an evaluation of the **inferences** results, with the classification metrics *Accuracy* and *F1 score*. These experiments were performed with the test set of the adapted EMOTIC database (which represents the 20% of the samples), which uses the AP and mAP as the standard to report the results (Kosti et al., 2017). Regarding the training phase, the descend gradient algorithm was used with a learning rate of 0.001; because classification is a multi-class and multi-label problem, the *Binary Cross-entropy* loss function with logistic was used. The training phase was performed in 32 epochs.

Single-modal Results Comparison. Results from each modality (body, contextual, facial, and postural) in terms of every emotion and in average are shown in Table 2. Concerning the results of AP metrics for each CG, the best result by modality varies. For Joy, the postural modality achieved a score of 82.22; however, this modality has too small values for other groups of emotions, not even reaching 5.0 score in most of them. The body modality achieves the best scores in Trust (71.43), Fear (17.29), Surprise (23.04), Sadness (14.85), and Anticipation (88.46). For Disgust, the contextual modality leads with 31.89; and for Anger, the facial modality has the highest AP score with 15.06. Regarding the mAP metric, the body modality leads with a score of 42.54, the contextual one is close with 42.31. Besides, the facial modality scores 38.72 and the postural modality 14.30, the worst one.

This evaluation shows that not all modalities perform well with data in the wild, where the presence of the face or the entire clear posture are not guaranteed. Thus, the performance of facial and postural modalities is affected, they are below the corporal modality in 3.82 units and 28.24 units, respectively. The postural results are not surprising because, in wild data, postural was expected to be the least expressive emotional data source. Likewise, this modality could improve its results using videos because the DGCNN model is better when using this type of data.

Table 3: Comparisons of the AP metric results between the proposal and state-of-the-art-methods.

Cat. Group	EMOTIC Model		EmotiCon		Ours
	Max	Mean	Max	Mean	
Anger	14.97	11.50	21.92	18.75	17.74
Anticip.	87.53	73.09	91.12	81.62	89.10
Disgust	21.32	13.12	43.12	24.55	31.94
Fear	16.89	11.40	23.65	18.92	18.15
Joy	77.16	46.06	83.26	60.60	82.51
Sadness	19.66	14.18	26.39	17.83	18.11
Surprise	29.63	24.22	35.12	26.25	22.45
Trust	78.35	36.93	68.65	42.18	72.45
mAP	43.19	28.81	49.15	36.34	44.06

Fusion Results Comparison. The entire proposed method was compared with the methods in the state-of-the-art in EMOTIC benchmark, EMOTIC model (Kosti et al., 2019) and EmotiCon (Mittal et al., 2020). Due to our adequacy of the dataset, the comparison of our results and those reported by the other methods was carried out with the maximum and the average value of each CG – e.g., the Trust group is represented by the maximum and average values of those reported for the emotions Confidence, Esteem, and Sympathy in EMOTIC model and EmotiCon. These results are reported in Table 3.

According to maximum AP scores (Max), our proposal achieves similar results to EMOTIC Model and EmotiCon in most emotions; however, concerning the mAP score, our model scores 44.06, below the score of EmotiCon 49.15. Compared with the EMOTIC Model, the proposed method surpasses the score in the groups of Joy, Fear, Disgust, Anger, and Anticipation; although, on mAP, our proposal exceeds it by almost one unit. For Trust group, the proposal overcomes the score of EmotiCon, but both are below EMOTIC Model.

Observing the average values (Mean), our proposal obtains higher mAP scores than EMOTIC and EmotiCon models: 44.06 vs. 28.81 and 36.34, respectively. With the Joy group, there is a score difference of almost 20, comparing our method to EmotiCon. Only in the Surprise group, the proposal results are below both models; even, it is below the achieved by only the body modality. The low performance of the state-of-the-art methods, with average values, is explained by the unbalance in the dataset since emotions with low scores lower the group average. Moreover, in the literature it is not specified whether, for example, an oversampling technique was used in the training phase to solve the unbalance.

The last comparison is among the fusion meth-

Table 4: Comparisons of the AP metric results between the proposed EmbraceNet+ and other fusion methods. I=intermediate data, F=final outputs.

Cat. Group	EmbraceNet (I)	EmbraceNet (F)	Concat +MLP	EmbraceNet+	Without F	Without WS
Anger	16.54	16.52	15.29	17.74	16.26	16.62
Anticipation	89.36	88.63	88.24	89.10	88.97	88.65
Disgust	31.52	30.20	28.05	31.94	29.45	31.32
Fear	17.42	18.23	17.46	18.15	17.66	17.37
Joy	82.25	81.56	80.13	82.51	81.08	82.00
Sadness	16.46	15.78	15.36	18.11	16.13	16.62
Surprise	22.33	22.48	21.76	22.45	22.65	22.40
Trust	73.19	73.30	71.12	72.45	72.88	71.88
mAP	43.63	43.34	42.18	44.06	43.13	43.36

ods with the same metrics, taking the two ways of using EmbraceNet (the one that uses (I)ntermediate data and the one that use the (F)inal outputs), a concatenation plus an MLP (Concat +MLP), and the EmbraceNet+ (see Table 4).

To notice the influence of adding other fusion methods, we also compare the EmbraceNet+ results without an extra Concatenation (EmbraceNet+ without F) and without a WS (EmbraceNet+ without WS) as the additional fusion methods. At the level of CGs, only with Fear and Anticipation, the EmbraceNet+ does not achieve the highest AP score, but the EmbraceNet (F) with 18.23 and the EmbraceNet (I) with 89.36, respectively. In any other emotion, EmbraceNet+ has the best results. For Trust and Surprise, the EmbraceNet+ without including the concatenation (Without F) achieves better results, with 72.88 and 22.65, respectively.

Our proposal, EmbraceNet+, achieves a slight improvement in mAP score over both EmbraceNets, with a difference of 0.43 with the EmbraceNet (I) and 0.72 with the EmbraceNet (F). The Concat +MLP does not achieve the highest result for any emotion; then, this method is not enough to solve the recognition of emotions as it was presented. If any of the additional fusion methods are removed, our EmbraceNet+ does not outperform EmbraceNets, therefore, adding these fusion strategies (Concatenation and WS) provides relevant data for better results. On the other hand, the modifications on the docking layers do not have the expected impact and not work as an additional feature extractor. This may be because the difference against normal EmbraceNet and normal docking layers, is short. However, it still maintains the correlation between modalities and makes the results of the docking layers more robust without falling into overfitting thanks to the dropout layer.

Evaluation of Inferences. As explain previously, the final inferences are binary vectors representing the presence or absence of the emotions and can be evaluated with classification metrics, such as accuracy and F1 score. These inferences were obtained from test and validation datasets; however, the results obtained are quite similar. Results with the test-set are 0.56% better than results obtained with the validation set, on average accuracy. Thus, we only show, in Table 5, the results with the test-set. The accuracy indicates a good performance, reaching up to 86% in Anger and Sadness groups and the lower in the Trust group with 62%, and 73% in average; the accuracy demonstrates that our model can estimate in a good way the presence and the absence of emotions. Results according to F1 score are like those of AP, because both are based on the precision and recall.

Table 5: Classification results in accuracy and F1 metrics.

Cat. Group	Accuracy	F1 score
Anger	0.8667	0.2005
Anticipation	0.7792	0.8686
Disgust	0.6626	0.3268
Fear	0.7375	0.1781
Joy	0.6881	0.7971
Sadness	0.8627	0.2206
Surprise	0.6994	0.2339
Trust	0.6223	0.6694
Average	0.7398	0.4369

We also measure the performance of our approach in terms of total time. In average, it is able to process one image and estimate the emotions of one person in **0.182 sec**, working with a video card Nvidia Tesla K80. This time considers the person detection in the image (0.055 sec), the proper data acquisition (0.088 sec), and the whole processing to achieve a prediction (0.038 sec). This result of the time execution provides the opportunity to implement the proposal in real-time or near-real-time applications, expanding the domains of applicability.

General Discussion. All results demonstrate the importance of employing a multi-modal approach to study emotional states in wild data since single modality methods remain below multi-modal ones, with AP and mAP metrics. Furthermore, techniques such as oversampling and clustering help preventing poor performance due to imbalance; however, they are not enough in some cases: scores of emotions with fewer samples remain low.

On the other hand, EmbraceNet+ is applicable in almost all multi-modal approaches to perform classification or recognition in any domain. Likewise, the proposed emotion recognition method can be implemented in other scenarios that require knowing the emotional state of people. These advantages of our proposed method for emotion recognition are supported by the implementation and usage of an ontology; the flexibility inherited from the EmbraceNet fusion method and the good performance obtained with EmbraceNet+, suggest that this novel fusion method is applicable in real-time or near real-time applications.

6 CONCLUSIONS

We present a multi-modal method to recognize emotions from images, whose values are stored in an emotion ontology, called EMONTO.

The multi-modal method uses deep learning models, considering four types of data: face expression, posture skeleton graph, features of body, and environmental context. For the fusion of modalities, we proposed EmbraceNet+, an architecture that integrates three fusion methods: the naive concatenation, the WS, and the EmbraceNet. Our method achieves similar and competitive results compared to state-of-the-art methods for the EMOTIC and EmotiCon benchmarks. Likewise, EmbraceNet+ can be used to merge any multi-modal deep learning method. We also showed how to instantiate EMONTO, using the emotions from the multi-modal method. EMONTO contains classes that allows integrating other specific domain ontologies, giving our approach the ability to be applied in multiple scenarios. Our proposed method offers a wide range of possibilities for scholarly research, as accurate connections of the kind can be used for the design and implementation of smart applications that exploit semantic web resources, in real-time or near real-time. Accordingly, we plan to apply our proposal in real scenarios such as museums as cases of study.

In future works, we also would explore other visual data, such as the interactions and proximity in similar ways (Mittal et al., 2020); additionally, we plan to explore other ways to use the EmbraceNet and another type of feature vector (Choi and Lee, 2019b).

ACKNOWLEDGEMENT

This research was supported by the FONDO NACIONAL DE DESARROLLO CIENTÍFICO, TECNOLÓGICO Y DE INNOVACIÓN TECNOLÓGICA - FONDECYT as executing entity of CONCYTEC under grant agreement no. 01-2019-FONDECYT-BM-INC.INV in the project RUTAS: Robots for Urban Tourism, Autonomous and Semantic web based.

REFERENCES

Ali, F., Kwak, D., Khan, P., Islam, S. R., Kim, K. H., and Kwak, K. (2017). Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. *Transportation Research Part C: Emerging Technologies*, 77:33 – 48.

Bertola, F. and Patti, V. (2016). Ontology-based affective models to organize artworks in the social semantic web. *Information Processing & Management*, 52(1):139–162.

Caruccio, L., Polese, G., Tortora, G., and Iannone, D. (2019). Edcar: A knowledge representation framework to enhance automatic video surveillance. *Expert Systems with Applications*, 131:190–207.

Cavaliere, D. and Senatore, S. (2019). Emotional concept extraction through ontology-enhanced classification. In *Research Conf. on Metadata and Semantics Research*, pages 52–63. Springer.

Chen, J., Hu, B., Moore, P., and Zhang, X. (2016). Ontology-based model for mining user's emotions on the wisdom web. In *Wisdom Web of Things*, pages 121–153. Springer.

Choi, J.-H. and Lee, J.-S. (2019a). Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270.

Choi, J.-H. and Lee, J.-S. (2019b). Embracenet for activity: A deep multimodal fusion architecture for activity recognition. In *Adjunct Proc. of the ACM Int. Joint Conf. on Pervasive and Ubiquitous Computing and Proc. of the ACM Int. Symp. on Wearable Computers*, pages 693–698.

Chumkamon, S., Hayashi, E., and Koike, M. (2016). Intelligent emotion and behavior based on topological consciousness and adaptive resonance theory in a companion robot. *Biologically Inspired Cognitive Architectures*, 18:51–67.

Cowen, A. S. and Keltner, D. (2020). What the face displays: Mapping 28 emotions conveyed by naturalistic expression. *American Psychologist*, 75(3):349.

Deng, J., Guo, J., Yuxiang, Z., Yu, J., Kotsia, I., and Zafeiriou, S. (2019). Retinaface: Single-stage dense face localisation in the wild. In *arxiv*.

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, pages 1–18.

Dragoni, M., Poria, S., and Cambria, E. (2018). Ontosenticnet: A commonsense ontology for sentiment analysis. *IEEE Intelligent Systems*, 33(3):77–85.

Egger, J., Pepe, A., Gsaxner, C., and Li, J. (2020). Deep learning–a first meta-survey of selected reviews across scientific disciplines and their research impact. *arXiv preprint arXiv:2011.08184*.

Farhadi, A. and Redmon, J. (2018). Yolov3: An incremental improvement. *Computer Vision and Pattern Recognition*.

Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2019). Attention branch network: Learning of attention mechanism for visual explanation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 10705–10714.

Garay-Vitoria, N., Cearreta, I., and Larraza-Mendiluze, E. (2019). Application of an ontology-based platform for developing affective interaction systems. *IEEE Access*, 7:40503–40515.

Gilda, S., Zafar, H., Soni, C., and Waghurdekar, K. (2017). Smart music player integrating facial emotion recognition and music mood recommendation. In *Int. Conf. on Wireless Communications, Signal Processing and Networking*, pages 154–158. IEEE.

Grassi, M. (2009). Developing heo human emotions ontology. In Fierrez, J., Ortega-Garcia, J., Esposito, A., Drygajlo, A., and Faundez-Zanuy, M., editors, *Biometric ID Management and Multimodal Communication*, pages 244–251, Berlin, Heidelberg. Springer Berlin Heidelberg.

Graterol, W., Diaz-Amado, J., Cardinale, Y., Dongo, I., Lopes-Silva, E., and Santos-Libarino, C. (2021). Emotion detection for social robots based on nlp transformers and an emotion ontology. *Sensors*, 21(4).

Katifori, A., Golemati, M., Vassilakis, C., Lepouras, G., and Halatsis, C. (2007). Creating an ontology for the user profile: Method and applications. pages 407–412.

Kaur, R. and Kautish, S. (2019). Multimodal sentiment analysis: A survey and comparison. *Int. Jrnl. of Service Science, Management, Engineering, and Technology*, 10(2):38–58.

Knapp, M. L., Hall, J. A., and Horgan, T. G. (2013). *Nonverbal communication in human interaction*, chapter 1: "Nonverbal Communication: Basic Perspectives". Cengage Learning, Boston, MA.

Kosti, R., Alvarez, J., Recasens, A., and Lapedriza, A. (2019). Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kosti, R., Alvarez, J. M., Recasens, A., and Lapedriza, A. (2017). Emotion recognition in context. In *IEEE Conf. on Computer Vision and Pattern Recognition*.

Lee, J., Kim, S., Kim, S., Park, J., and Sohn, K. (2019). Context-aware emotion recognition networks. In *The IEEE Int. Conf. on Computer Vision*.

Lhommet, M., Marsella, S. C., Calvo, R., D'Mello, S., Gratch, J., and Kappas, A. (2015). *The Oxford Handbook of Affective Computing*, chapter "Expressing Emotion Through Posture and Gesture".

Lin, R., Amith, M. T., Liang, C., Duan, R., Chen, Y., and Tao, C. (2018a). Visualized emotion ontology: A model for representing visual cues of emotions. *BMC Medical Informatics and Decision Making*, 18.

Lin, R., Liang, C., Duan, R., Chen, Y., Tao, C., et al. (2018b). Visualized emotion ontology: a model for representing visual cues of emotions. *BMC medical informatics and decision making*, 18(2):64.

Liu, K., Li, Y., Xu, N., and Natarajan, P. (2018). Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*.

Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 14234–14243.

Noroozi, F., Kaminska, D., Corneanu, C. A., Sapinski, T., Escalera, S., and Anbarjafari, G. (2018). Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conf.*

Perez-Gaspar, L.-A., Caballero-Morales, S.-O., and Trujillo-Romero, F. (2016). Multimodal emotion recognition with evolutionary computation for human-robot interaction. *Expert Systems with Applications*, 66:42–61.

Pinto-De la Gala, A., Cardinale, Y., Dongo, I., and Ticona-Herrera, R. (2021). Towards an ontology for urban tourism. In *Proc. of the 36th Annual ACM Symp. on Applied Computing*, SAC '21, New York, NY, USA. ACM.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019). Skeleton-based action recognition with directed graph neural networks. In *The IEEE Conf. on Computer Vision and Pattern Recognition*.

Soleymania, M., Garcia, D., Jouc, B., Schuller, B., Chang, S.-F., and Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *The IEEE Conf. on Computer Vision and Pattern Recognition*.

Zadeh, M. M. T., Imani, M., and Majidi, B. (2019). Fast facial emotion recognition using convolutional neural networks and gabor filters. In *5th Conf. on Knowledge Based Engineering and Innovation*, pages 577–581. IEEE.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Zhang, S.-F., Zhai, J.-H., Xie, B.-J., Zhan, Y., and Wang, X. (2019). Multimodal representation learning: Advances, trends and challenges. In *Int. Conf. on Machine Learning and Cybernetics*, pages 1–6. IEEE.

Zhang, X., Hu, B., Chen, J., and Moore, P. (2013). Ontology-based context modeling for emotion recognition in an intelligent web. *World Wide Web*, 16(4):497–513.