Emotion Recognition in Context

Ronak Kosti*, Jose M. Alvarez†, Adria Recasens‡, Agata Lapedriza*
Universitat Oberta de Catalunya*
Data61 / CSIRO†
Massachusetts Institute of Technology‡
{rkosti, alapedriza}@uoc.edu*, jalvarez@cvc.uab.es†, recasens@mit.edu‡

Abstract

Understanding what a person is experiencing from her frame of reference is essential in our everyday life. For this reason, one can think that machines with this type of ability would interact better with people. However, there are no current systems capable of understanding in detail people's emotional states. Previous research on computer vision to recognize emotions has mainly focused on analyzing the facial expression, usually classifying it into the 6 basic emotions [11]. However, the context plays an important role in emotion perception, and when the context is incorporated, we can infer more emotional states. In this paper we present the "Emotions in Context Database" (EMOTIC), a dataset of images containing people in context in non-controlled environments. In these images, people are annotated with 26 emotional categories and also with the continuous dimensions valence, arousal, and dominance [21]. With the EMOTIC dataset, we trained a Convolutional Neural Network model that jointly analyses the person and the whole scene to recognize rich information about emotional states. With this, we show the importance of considering the context for recognizing people's emotions in images, and provide a benchmark in the task of emotion recognition in visual context.

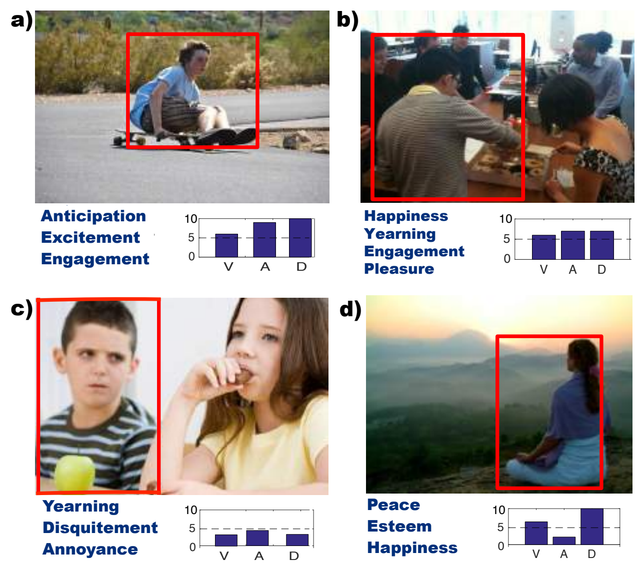

Figure 1: How do these people feel?

1. Introduction

Understanding how people feel plays a crucial role in social interaction. This capacity is necessary to perceive, anticipate and respond with care to people's reactions. We are remarkably skilled at this task and we regularly make guesses about people's emotions in our everyday life. Particularly, when we observe someone, we can estimate a lot of information about that person's emotional state, even without any additional knowledge about this person. As an example, take a look at the images of Figure 1. Let us put ourselves in these people's situations and try to estimate what they feel. In Figure 1.a we can recognize an emotional state of *anticipation*, since this person is constantly looking at the road to correctly adapt his trajectory. We can also recognize that this person feels *excitement* and that he is *engaged* or absorbed with the activity he is performing. We can also say that the overall emotion that he is feeling is *positive*, he is *active* and he seems confident with the activity he is performing, so he is *in control* of the situation. Similar detailed estimations can be made about the people marked with a red rectangle in the other images of Figure 1.

Recognizing people's emotional states from images is an active area of research among the computer vision community. Section 2 describes some of the recent works in this topic. Overall, in the last years we observe an impressive progress in recognizing the 6 basic emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*) from facial expression. Some interesting efforts have been also in the understanding of body language and in the use body pose features to recognize some specific emotional states. However, in this previous research on emotion recognition, the context of the subject is usually ignored.

Some works in psychology show evidences on the importance of context in the perception of emotions [3]. In most of the cases, when we analyze a wider view instead of focusing on the person, we can recognize additional affective information that can not be recognized if the context is not taken into account. For example, in Figure 1(c), we can see that the boy feels *annoyance* because he has to eat an apple while the girl next to him has chocolate, which is something that he feels *yearning* (strong desire) for. The presence of the girl, the apple and the chocolate are necessary clues to understand well what he indicates with his facial expression.

In fact, if we consider the context, we can make reasonable guesses about emotional states even when the face of the person not visible, as illustrated in Figures 1.b and 1.d. The person in the red rectangle of Figure 1.b is picking a doughnut and he probably feels *yearning* to eat it. He is participating in a social event with his colleagues, showing *engagement*. He is feeling *pleasure* eating the doughnuts and *happiness* for the relaxed break along with other people. In Figure 1.d, the person is admiring the beautiful landscape with *esteem*. She seems to be enjoying the moment (*happiness*), and she seems calmed and relaxed (*peace*). We do not know exactly what is on the people's minds, but we are able to reasonably extract relevant affective information just by looking at them in their situations.

This paper addresses the problem of recognizing emotional states of people in context. The first contribution of our work is the *Emotions in Context Database* (EMOTIC), which is described in Section 2. The EMOTIC database is a collection of images with people in their context, annotated according to the emotional states that an observer can estimate from the whole situation. Specifically, images are annotated with two complementary systems: (*1*) an extended list of 26 affective categories that we collected, and (*2*) the common continuous dimensions $Valence$, $Arousal$, and $Dominance$ [21]. The combination of these two emotion representation approaches produces a detailed model that gets closer to the richness of emotional states that humans can perceive [2].

Using the EMOTIC database, we test a Convolutional Neural Network (CNN) model for recognizing emotions in context. Section 4 describes the model, while Section 5 presents our experiments. From our results, we make two interesting conclusions. First, we see that the context contributes relevant information for emotional states recognition. Second, we observed that combining categories and continuous dimensions during the training results in a more robust system for recognizing emotional states.

2. Related work

Most of the research in computer vision to recognize emotional states in people is contextualized in facial expression analysis (e.g.,[4, 13]). We find a large variety of methods developed to recognize the 6 basic emotions defined by the psychologists Ekman and Friesen [11]. Some of these methods are based on the *Facial Action Coding System* [15, 29]. This system uses a set of specific localized movements of the face, called *Action Units*, to encode the facial expression. These Action Units can be recognized from geometric-based features and/or appearance features extracted from face images [23, 19, 12]. Recent works for emotion recognition based on facial expression use CNNs to recognize the emotions and/or the Action Units [4].

Instead of recognizing emotion categories, some recent works on facial expression [28] use the continuous dimensions of the *VAD Emotional State Model* [21] to represent emotions. The VAD model describes emotions using 3 numerical dimensions: ***Valence*** (V), that measures how positive or pleasant an emotion is, ranging from *negative* to *positive*; ***Arousal*** (A), that measures the agitation level of the person, ranging from *non-active / in calm* to *agitated / ready to act*; and ***Dominance*** (D) that measures the control level of the situation by the person, ranging from *submissive / non-control* to *dominant / in-control*. On the other hand, Du et al. [10] proposed a set of 21 facial emotion categories, defined as different combinations of the basic emotions, like 'happily surprised' or 'happily disgusted'. This categorization gives more detail about the expressed emotion.

Although most of the works in recognizing emotions are focused on face analysis, there are a few works in computer vision that address emotion recognition using other visual clues apart from the face. For instance, some works [22] consider the location of shoulders as additional information to the face features to recognize basic emotions. More generally, Schindler et al. [27] used the body pose to recognize the 6 basic emotions, performing experiments on a small dataset of non-spontaneous poses acquired under controlled conditions.

In the recent years we also observed a significant emergence of affective datasets to recognize people's emotions. The studies [17, 18] establish the relationship between affect and body posture using as ground truth the base-rate of human observers. The data consists of a spontaneous subset acquired under a restrictive setting (people playing Wii games). In EMOTIW challenge [7], AFEW database [8] focuses on emotion recognition in video frames taken from movies and TV shows, while the HAPPEI database [9] addresses the problem of group level emotion estimation. In this work we can see a first attempt to the use context for the problem of predicting happiness in groups of people. Finally, the MSCOCO dataset has been recently annotated with object attributes [24], including some feeling categories for people, such as *happy* or *curious*. These attributes show some overlap with the categories that we define in this paper. However, the COCO attributes are not intended to be exhaustive for emotion recognition, and not all the people in the dataset are annotated with affect attributes.

3. Emotions in Context Database (EMOTIC)

The EMOTIC database is composed of images from MSCOCO [20], Ade20k [31] and images that we manually downloaded from Google search engine. For the latter, we used as queries different words related with feelings and emotions. This combination of images results in a challenging collection with images of people performing different activities, in different places, showing a huge variety of emotional states. The database contains a total number of $18,316$ images having $23,788$ annotated people.

Figure 1 shows examples of images in the database along with their annotations. As shown, our database combines two different emotion representation formats:

Discrete Categories: a set of 26 emotional categories, that cover a wide range of emotional states. The categories are listed and defined in table 1, while Figure 2 shows, per each category, examples of people labeled with the corresponding category.

Continuous Dimensions: the three emotional dimensions of the VAD Emotional State Model [21]. The continuous dimensions annotations in the database are in a $1-10$ scale. Figure 3 shows examples of people with different levels of each one of these three dimensions.

In order to define the proposed set of emotional categories of Table 1, we collected an extensive vocabulary of affective states, composed of more than 400 words. We grouped these words using dictionaries (definitions and synonyms) and books on psychology and affective computing [14, 25]. After this process we obtained 26 groups of words. We named each group with a single affective word representing all the words in the group, and these became our final list of 26 categories. While grouping the vocabulary, we wanted to fulfill 2 conditions: (*i*) *Disjointness*: given any category pair, $\{c_1, c_2\}$, we could always find an example of image where just one of the categories apply (and not the other); and (*ii*) *Visual separability*: two words were assigned to the same group in case we find, qualitatively, that the two could not be visually separable under the conditions of our database. For example, the category *excitement* includes the subcategories "enthusiastic, stimulated, and energetic". Each of these three words have a specific and different meaning, but it is very difficult to separate one from another just after seeing a single image. In our list of categories we decided to avoid the *neutral* category since we think that, in general, at least one category applies, even though it could apply just with low intensity.

Notice that our final list of categories (see Table 1) includes the 6 basic emotions (categories $7, 10, 18, 20, 22, 24$) [26], however, category 18 ($Aversion$) is a generic version of the basic emotion $disgust$. The $Aversion$ category groups the subcategories $disgust, dislike, repulsion,$ and $hate$, which are difficult to separate visually.

1. Peace: well being and relaxed; no worry; having positive thoughts or sensations; satisfied
2. Affection: fond feelings; love; tenderness
3. Esteem: feelings of favorable opinion or judgment; respect; admiration; gratefulness
4. Anticipation: state of looking forward; hoping on or getting prepared for possible future events
5. Engagement: paying attention to something; absorbed into something; curious; interested
6. Confidence: feeling of being certain; conviction that an outcome will be favorable; encouraged; proud
7. Happiness: feeling delighted; feeling enjoyment or amusement
8. Pleasure: feeling of delight in the senses
9. Excitement: feeling enthusiasm; stimulated; energetic
10. Surprise: sudden discovery of something unexpected
11. Sympathy: state of sharing others emotions, goals or troubles; supportive; compassionate
12. Doubt/Confusion: difficulty to understand or decide; thinking about different options
13. Disconnection: feeling not interested in the main event of the surrounding; indifferent; bored; distracted
14. Fatigue: weariness; tiredness; sleepy
15. Embarrassment: feeling ashamed or guilty
16. Yearning: strong desire to have something; jealous; envious; lust
17. Disapproval: feeling that something is wrong or reprehensible; contempt; hostile
18. Aversion: feeling disgust, dislike, repulsion; feeling hate
19. Annoyance: bothered by something or someone; irritated; impatient; frustrated
20. Anger: intense displeasure or rage; furious; resentful
21. Sensitivity: feeling of being physically or emotionally wounded; feeling delicate or vulnerable
22. Sadness: feeling unhappy, sorrow, disappointed, or discouraged
23. Disquietment: nervous; worried; upset; anxious; tense; pressured; alarmed
24. Fear: feeling suspicious or afraid of danger, threat, evil or pain; horror
25. Pain: physical suffering
26. Suffering: psychological or emotional pain; distressed; anguished

Table 1: Proposed emotional categories with definitions.

3.1. Image Annotation

We designed Amazon Mechanical Turk (AMT) interfaces to annotate feelings according to the specified categories and continuous dimensions. In the case of categories, we show an image with a person marked, and we ask the workers to select all the feeling categories that apply for that person in that situation. In the case of the continuous dimensions, we ask the workers to select the proper score per each one of the dimensions. Furthermore workers also annotated the gender (male/female) as well as the age range (child, teenager, adult) of each person.

We followed two strategies to ensure the quality of the annotations. First, workers needed to pass a qualification

Figure 2: Visual examples of the 26 feeling categories defined in Table 1. Per each category we show two images where the person marked with the red bounding box has been annotated with the corresponding category.

Figure 3: Examples of people with difference scores of Valence (row 1), Arousal (row 2) and Dominance (row 3).

task before annotating images from the dataset. This task consisted of a standard Emotional Quotient test [16] plus two image annotations. Once the workers passed this qualification task they were allowed to annotate images of the database. Second, while workers were annotating images, we monitored their performance by adding 2 control images for every 18 images.

The final dataset is split in training (70%), validation (10%), and testing (20%) sets. The testing set was annotated by 3 different workers, in order to check the consistency of annotations among different people. Although the guess that one can make about other's emotional states is subjective, we observed a significant degree of agreement among different annotators. In particular, if a worker selects a category, the probability that he is agreeing with at least one of the other two other workers in selecting this category is 23.97%. We also computed the *Fleiss' kappa* for each image, obtaining a mean value of 0.27. More than 50% of the images have $\kappa > 0.33$. These statistics show a reasonable degree of agreement in the discrete categories (notice that random annotations will produce a 0 kappa value). Regarding to the continuous dimensions, the standard deviation among the different workers is, in average, 1.41, 0.70 and 2.12 for valence, arousal and dominance respectively. This means that the higher dispersion is shown for dominance, where the annotations are around ± 2 from the average value.

3.2. Database Statistics

Of the $23,788$ annotated people, 66% are males and 34% are females. Their ages are distributed as follows: 11% children, 11% teenagers, and 78% adults. Figure 4.a shows the number of people for each of the categories, Figures 4.b, 4.c and 4.d show the number of people for valence, arousal and dominance continuous dimensions for each score respectively.

We observed that our data shows interesting patterns of category co-occurrences. For example, after computing conditional probabilities, we could see that when a person feels *affection* it is very likely that she also feels *happiness*, or that when a person feels *anger* she is also likely to feel *annoyance*. More generally, we used k-means to cluster the category annotations and observed that some category groups appear frequently in our database. Some examples are $\{$*anticipation*, *engagement*, *confidence*$\}$, $\{$*affection*, *happiness*, *pleasure*$\}$, $\{$*doubt/confusion*, *disapproval*, *annoyance*$\}$, $\{$*yearning*, *annoyance*, *disquietment*$\}$.

Figure 5 displays, for each continuous dimension, the %

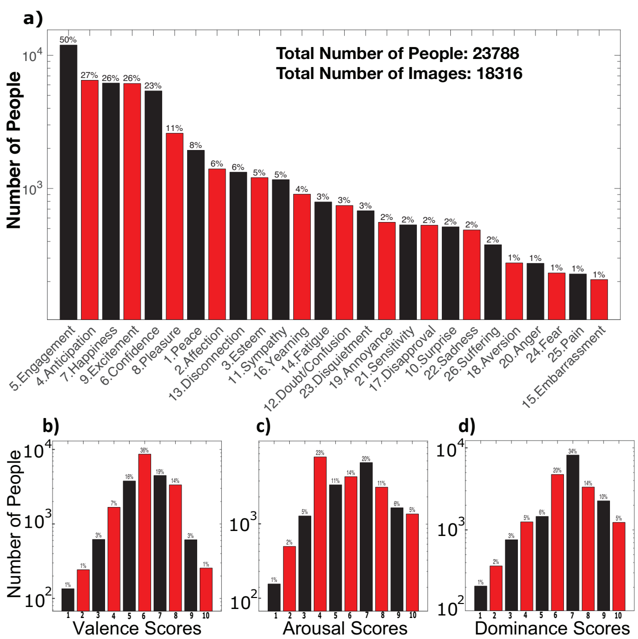

Figure 4: Dataset statistics. (a) Number of examples per feelings category; (b), (c), and (d) Number of examples per each of the scores in the continuous dimensions: (b) valence, (c) arousal, (d) dominance.

distribution of scores for each of the discrete categories. In each case, categories are sorted in increasing order according to the average of the corresponding dimension. Thus, Figure 5.a shows how the valence scores are distributed. The first categories (*suffering* and *pain*) are the ones with lowest Valence score in average, while the last ones, (*affection* and *happiness*), are the ones that have highest valence score on average. This makes sense, since *suffering* and *pain* are in general unpleasant feelings, while *confidence* and *excitement* are usually pleasant feelings. In Figure 5.b we can see the same type of information for the *arousal* dimension. Notice that *fatigue* and *sadness* are the categories that, have lowest *arousal* score in average, meaning that when these feelings are present, people are usually in a low level of agitation. On the other hand, *confidence* and *excitement* are the categories with highest *arousal* level. Finally, Figure 5.c shows the distribution of the *dominance* scores. The categories with lowest *dominance* level (people feeling they are not in control of the situation) are *suffering* and *pain*, while the highest dominance levels in average are shown with the categories *confidence* and *excitement*. We observe that these types of category sorting are consistent with our common sense knowledge. However, we also observe in these graphics that per each category we have a some relevant variability of the continuous dimension scores. This suggests that the information contributed

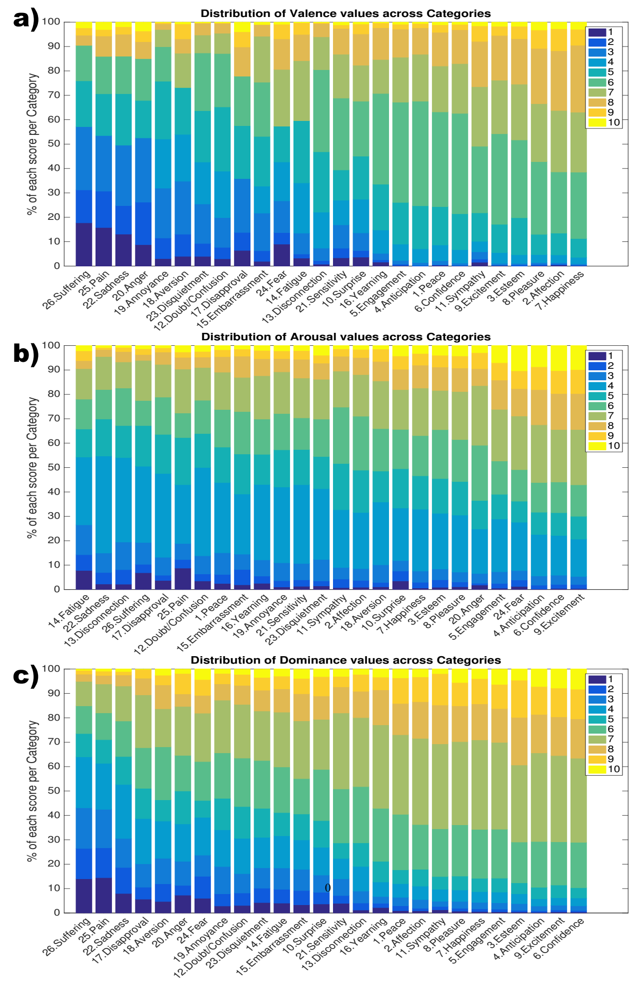

Figure 5: Per each continuous dimension, distribution of the scores across the different categories. Categories are sorted according to the average value of the corresponding category, from the lowest (left) to the highest (right).

by each type of annotation can be complementary and not redundant.

Finally, we want to highlight that a considerable percentage of the images in the EMOTIC Database do not have the face clearly visible, like in the examples shown in Figure 1.b and 1.b. After randomly selecting subsets of 300 images and counting how many people do not have the face visible, we estimate that this happens in 25% of the samples. Among the rest of the samples, a lot of them have significant partial occlusions in the face, or faces are shown in non-frontal views. For this reason, the task of estimating how the people feel in this database can not be approached with facial expression analysis, presenting us with a new challenging task.

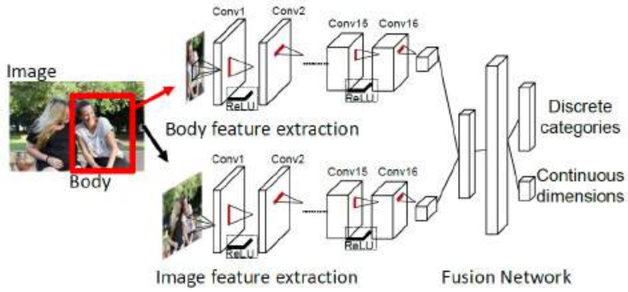

Figure 6: Proposed end-to-end model for emotion recognition in context. The model consists of two modules for extracting features and a fusion network for jointly estimating the discrete categories and the continuous dimensions.

4. Proposed CNN Model

We propose an end-to-end model, shown in Figure 6, for simultaneously estimating the discrete categories and the continuous dimensions. The architecture consists of three main modules: two feature extractors and a fusion module. The first module takes the region of the image comprising the person who's feelings are to be estimated and extracts its most relevant features. The second module takes as input the entire image and extracts global features for providing the necessary contextual support. Finally, the third module is a fusion network that takes as input the image and body features and estimates the discrete categories and the continuous dimensions. The parameters of the three modules are learned jointly.

Each feature extraction module is designed as a truncated version of the low-rank filter convolutional neural network proposed in [1]. The main advantage of this network is providing competitive accuracy while maintaining the number of parameters and computational complexity to a minimum. The original network consists of 16 convolutional layers with 1-dimensional kernels, effectively modeling 8 layers using 2-dimensional kernels. Then, a fully connected layer directly connected to the softmax layer. Our truncated version removes the fully connected layer and outputs features from the activation map of the last convolutional layer. The rationale behind this is preserving the localization of different parts of the image which is relevant for the task at hand.

Features extracted from these two modules are combined using a separate fusion network. This fusion module first uses a global average pooling layer on each feature map to reduces the number of features, and then, a first fully connected layer acts as a dimensionality reduction layer for the set of concatenated pooled features. The output of this layer is a 256 dimensional vector. Then, we include a large fully connected layer to allow the training process learning independent representations for each task [5]. This layer is split into two separate branches: one for the continuous dimensions and the other for the discrete categories containing 26 and 3 neurons respectively. Batch normalization and rectifier linear units are added after each convolutional layer.

The parameters of the three modules are learned jointly using stochastic gradient descent with momentum. We set the batch size to 52 which corresponds to twice the number of discrete categories in the dataset. We use uniform sampling per category to have at least one instance of each discrete category in each batch. Empirically, using this approach, we obtained better results compared to shuffling the training set. The overall loss for training the model is defined as a weighted combination of two individual losses: $L_{comb} = \lambda_{disc} L_{disc} + \lambda_{cont} L_{cont}$. The parameter $\lambda_{disc,cont}$ weights the importance of the each loss and, L_{disc} and L_{cont} represent the loss corresponding to the tasks of learning discrete categories and learning the continuous dimensions respectively.

Discrete dimensions: We formulate this multiclass-multilabel problem as a regression problem using a weighted Euclidean loss to compensate for the class imbalance existing in the dataset. We empirically found this loss to be more effective than using Kullback-Leibler divergence or a multi-class multi-classification hinge loss. More precisely, this loss is defined as follows,

$$L_{disc} = \frac{1}{N} \sum_{i=1}^{N} w_i (\hat{y}_i^{disc} - y_i^{disc})^2 \qquad (1)$$

where N is the number of categories (N=26 in our case), $\hat{y}_i^{disc}$ is the estimated output for the i-th category and y_i^{disc} is the ground-truth label. The parameter w_i is the weight assigned to each category. Weight values are defined as $w_i = \frac{1}{ln(c+p_i)}$, where p_i is the probability of the i-th category and c is a parameter to control the range of valid values for w_i. Using this weighting scheme the values of w_i are bounded as the number of instances of a category approach to 0. This is particularly relevant in our case as we set the weights based on the occurrence of each category in every batch. Empirically we obtained better results using this approach compared to setting the weights based on the complete dataset at once.

Continuous dimensions: We formulate this task as a regression problem using the Euclidean loss. In this case, we consider an error margin to compensate fluctuations in the labeling process due to multiple workers labeling the data using a subjective, and not normalized, assessment. The loss for the continuous dimensions is defined as follows,

$$L_{cont} = \frac{1}{\sharp \mathcal{C}} \sum_{k \in \mathcal{C}} v_k (\hat{y}_k^{cont} - y_k^{cont})^2 \qquad (2)$$

where $\mathcal{C} = \{Valence, Arousal, Dominance\}$, $\hat{y}_k^{cont}$ and $\hat{y}_k^{cont}$ are the estimated output and the normalized ground-truth for the k-th dimension and $v_k = [0, 1]$ is a weight to represent the error margin. $v_k = 0$ if $|\hat{y}_k^{cont} - y_k^{cont}| < \theta$. Otherwise, $v_k = 1$. That is, there is no loss corresponding to estimations which error is smaller than θ and therefore these estimations do not take part during back-propagation.

We initialize the feature extraction modules using pretrained models from two different large scale classification datasets such as ImageNet [6] and Places [30]. ImageNet contains images of generic objects including person and therefore is a good option for understanding the contents of the image region comprising the target person. On the other hand, Places is a dataset specifically created for high level visual understanding tasks such as recognizing scene categories. Hence, pretraining the image feature extraction model using this dataset ensures providing global (high level) contextual support.

5. Experiments and Discussion

We trained different configurations of the CNN model showed in Figure 6, with different inputs and different loss functions and evaluated the models with the testing set. In all the cases, the parameters of the training are set using the validation set.

Table 2 shows the Average Precision (AP, area under the Precision Recall curves) obtained by the test set in the different categories. The first 3 columns are results obtained with a combined loss function (L_{comb}) , with CNN architectures that take as inputs just the body (B, first column), just the image (I, second columns), and the combined body and image ($B + I$, third column). We observe that for all of the categories, except *esteem*, the best result is obtained when both the body and the image are used as inputs. This shows that in the case of discrete category recognition, to consider both information sources is the best option. We note here that the results obtained using only the image (I) are, in general, worst amongst the three (B, I, $B + I$). This means that the image is helping in recognizing the emotions, but the scene itself does not provide enough information for the final recognition. This makes sense, since in the same scene we can have different people showing different emotions, although they are sharing most of their context. Another comparison shown in Table 2 is between the use of a combined loss function (L_{comb}) and the loss function trained only on the discrete categories (L_{disc}), considering $B + I$ as input. Notice that the performance is better in the case of the L_{comb}, demonstrating clearly that learning the continuous dimensions aids the learning of the emotional categories.

Table 3 shows the results for the continuous dimensions using error rates - *the difference (in average) between the true value and the regressed value*. Again, we note similar behavior as before: the best result is obtained with the combined loss function and using $B + I$ as inputs. However, the results are more uniform in the case of continuous dimensions.

Category	CNN Inputs and Loss			
	B	I	B + I	B + I
	L_{comb}			L_{disc}
1. Peace	20.63	20.43	**22.94**	20.03
2. Affection	21.98	17.74	**26.01**	20.04
3. Esteem	18.83	**19.31**	18.58	18.95
4. Anticipation	54.31	49.06	**58.99**	52.59
5. Engagement	82.17	78.48	**86.27**	80.48
6. Confidence	74.33	65.42	**81.09**	69.17
7. Happiness	54.78	49.32	**55.21**	52.81
8. Pleasure	48.65	45.38	48.65	**49.23**
9. Excitement	74.16	68.82	**78.54**	70.83
10. Surprise	21.95	19.71	**21.96**	20.92
11. Sympathy	11.68	11.30	**15.25**	11.11
12. Doubt/Confusion	33.49	33.25	**33.57**	33.16
13. Disconnection	18.03	16.93	**21.25**	16.25
14. Fatigue	9.53	7.30	**10.31**	7.67
15. Embarrassment	2.26	1.87	**3.08**	1.84
16. Yearning	8.69	7.88	**9.01**	8.42
17. Disapproval	12.32	6.60	**16.28**	10.04
18. Aversion	8.13	3.59	**9.56**	7.81
19. Annoyance	11.62	6.04	**16.39**	11.77
20. Anger	7.93	5.15	**11.29**	8.33
21. Sensitivity	5.86	4.94	**8.94**	4.91
22. Sadness	9.44	6.28	**19.29**	7.26
23. Disquietment	18.75	16.85	**20.13**	18.21
24. Fear	15.73	14.60	**16.44**	15.35
25. Pain	6.02	2.98	**10.00**	4.17
26. Suffering	10.06	5.35	**17.60**	7.42
Average	25.44	22.48	**28.33**	24.18

Table 2: Average precision obtained per each category for the different CNN input configurations: Body(**B**), Image(**I**), Body+Image(**B+I**).

Dimension	CNN Inputs and Loss			
	B	I	B + I	B + I
	L_{comb}			L_{cont}
Valence	**0.9**	**0.9**	**0.9**	1.0
Arousal	**1.1**	1.9	1.2	1.5
Dominance	1.0	**0.8**	0.9	**0.8**
Average	**1.0**	1.2	**1.0**	1.1

Table 3: Mean error rate obtained per each dimension for the different CNN input configurations: Body(**B**), Image(**I**), Body+Image(**B+I**).

Figure 8 shows another evaluation of our best model ($B + I$, L_{comb}). Using the validation set, we take as threshold, for each category, the value where *Precision* = *Recall*. We use the corresponding threshold to detect each category. After that, we computed the Jaccard coefficient for the recognized categories. That is, the number of detected categories that are also present in the ground truth divided-by the total number of active categories (detected + ground truth). We computed this Jaccard coefficient per all the images in the testing set, and they are shown in Figure 8.a. We observe that more than 25% of the samples in the testing

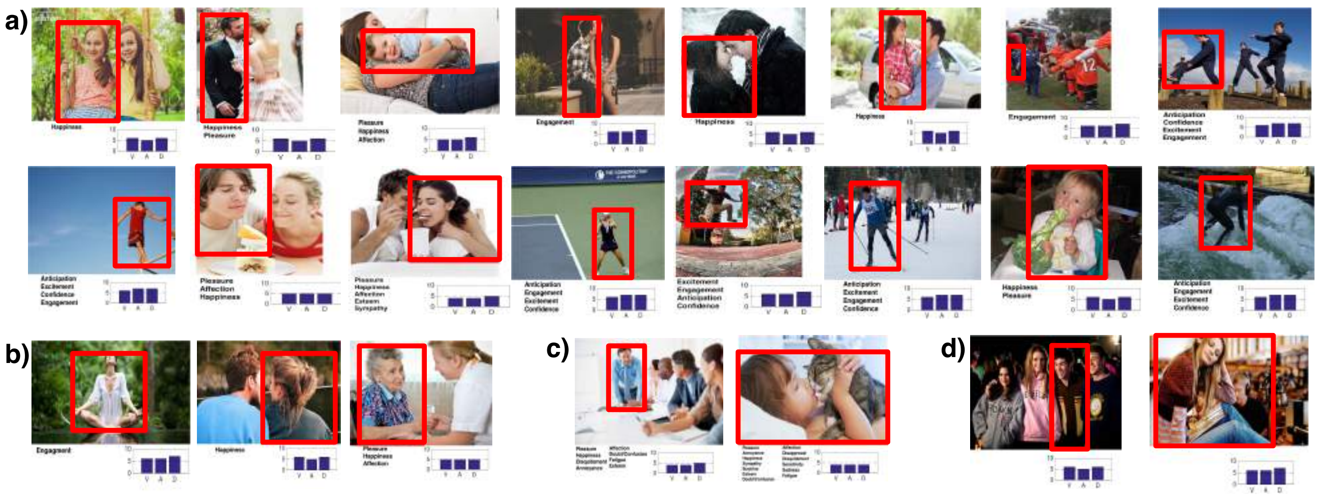

Figure 7: Results of emotion recognition in images of the testing set.

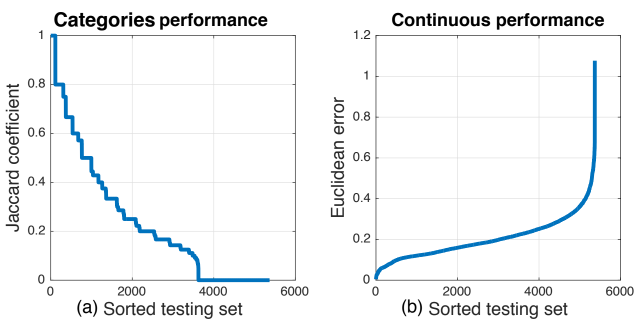

Figure 8: Results obtained in the testing set: (a) per each testing sample (sorted), Jaccard coefficient of the recognized discrete categories (b) per each testing sample (sorted), euclidean error in the estimation of the three continuous dimensions.

set have Jaccard coefficient above 0.4, which indicates that a significant number of categories in the ground truth were correctly retrieved. Figure 8.b shows the mean error rate obtained per each sample in the testing set when estimating the continuous dimensions. We observe that most of the examples have an error lower than 0.5.

Finally, Figure 7 shows some qualitative results. Images in Figure 7.a are randomly selected from those that have a Jaccard index for the categories recognition higher than 0.4 and an error on the continuous dimensions lower than 0.5. These could be considered as correct feeling estimations. The images in Figure 7.b are randomly selected across those that have a Jaccard index for the category recognition lower than 0.2. Regarding to the miss-classifications, we observed two common patterns when the Jaccard index is low: *(1)* in some images a high number of categories fire (examples shown in Figure 7.c), while *(2)* in some images none of the categories fire (examples shown in Figure 7.d). We observed, however, that even when the categories are not correctly recognized, the continuous dimensions are frequently well estimated. So, our method is still able to give reasonable information for these cases. Notice that, in general, our system is able to make significant guesses on emotional states even when the face of the person is not visible.

6. Conclusions

This paper addresses the problem of emotional state recognition in context. We present the EMOTIC database, a dataset of images in non-controlled environments containing people in context. The images are annotated according the people's apparent emotional states, combining 2 different types of annotations: the 26 emotional categories, proposed and described in this work, and the 3 standard continuous emotional dimensions (Valence, Arousal, and Dominance). We also proposed a CNN model for the task of emotion estimation in context. The model is based on state-of-the-art techniques for visual recognition and provides a benchmark on the proposed problem of estimating emotional states in context.

A technology capable of recognizing feelings in the same way as humans do has a lot of potential applications in human-computer interaction, human-assistive technologies and online education, among others.

Acknowledgments

This work has been partially supported by the *Ministerio de Economia, Industria y Competitividad (Spain)*, under the Grant Ref. TIN2015-66951-C2-2-R. The authors also thank NVIDIA for their generous hardware donations.

References

[1] J. Alvarez and L. Petersson. Decomposeme: Simplifying convnets for end-to-end learning. *CoRR*, abs/1606.05426, 2016.

[2] S. Baron-Cohen. Tead, the (2003) mind reading: The interactive guide to emotion.

[3] L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

[4] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR16), Las Vegas, NV, USA*, 2016.

[5] R. Caruana. *A Dozen Tricks with Multitask Learning*, pages 163–189. 2012.

[6] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016, pages 427–432, New York, NY, USA, 2016. ACM.

[8] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Acted facial expressions in the wild database. *Australian National University, Canberra, Australia, Technical Report TR-CS-11*, 2, 2011.

[9] A. Dhall, J. Joshi, I. Radwan, and R. Goecke. Finding happiest moments in a social context. In *Asian Conference on Computer Vision*, pages 613–626. Springer, 2012.

[10] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[11] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[12] S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing*, 24(1):189–204, 2015.

[13] S. Eleftheriadis, O. Rudovic, and M. Pantic. Joint facial action unit detection and feature fusion: A multi-conditional learning approach. *IEEE Transactions on Image Processing*, 25(12):5727–5742, 2016.

[14] E. G. Fernández-Abascal, B. García, M. Jiménez, M. Martín, and F. Domínguez. *Psicología de la emoción*. Editorial Universitaria Ramón Areces, 2010.

[15] E. Friesen and P. Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 1978.

[16] Y. Groen, A. B. M. Fuermaier, A. E. Den Heijer, O. Tucha, and M. Althaus. The empathy and systemizing quotient: The psychometric properties of the dutch version and a review of the cross-cultural stability. *Journal of Autism and Developmental Disorders*, 45(9):2848–2864, 2015.

[17] A. Kleinsmith and N. Bianchi-Berthouze. Recognizing affective dimensions from body posture. In *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*, ACII '07, pages 48–58, Berlin, Heidelberg, 2007. Springer-Verlag.

[18] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(4):1027–1038, Aug 2011.

[19] Z. Li, J.-i. Imai, and M. Kaneko. Facial-component-based bag of words and phog descriptor for facial expression recognition. In *SMC*, pages 1353–1358, 2009.

[20] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[21] A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*, 1995.

[22] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[23] M. Pantic and L. J. Rothkrantz. Expert system for automatic analysis of facial expressions. *Image and Vision Computing*, 18(11):881–905, 2000.

[24] G. Patterson and J. Hays. Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision*, pages 85–100. Springer, 2016.

[25] R. W. Picard and R. Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[26] J. Prinz. Which emotions are basic. *Emotion, evolution, and rationality*, 69:88, 2004.

[27] K. Schindler, L. Van Gool, and B. de Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural networks*, 21(9):1238–1246, 2008.

[28] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic. Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, 2016.

[29] J. F. C. W.-S. Chu, F. De la Torre. Selective transfer machine for personalized facial expression analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2015.

[31] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through ade20k dataset. 2016.