EMOTIC: Emotions in Context Dataset

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

Recognizing people's emotions from their frame of reference is very important in our everyday life. This capacity helps us to perceive or predict the subsequent actions of people, interact effectively with them and to be sympathetic and sensitive toward them. Hence, one should expect that a machine needs to have a similar capability of understanding people's feelings in order to correctly interact with humans. Current research on emotion recognition has focused on the analysis of facial expressions. However, recognizing emotions requires also understanding the scene in which a person is immersed. The unavailability of suitable data to study such a problem has made research in emotion recognition in context difficult. In this paper, we present the ***EMOTIC*** *database (from EMOTions In Context), a database of images with people in real environments, annotated with their apparent emotions. We defined an extended list of* 26 *emotion categories to annotate the images, and combined these annotations with three common continuous dimensions: Valence, Arousal, and Dominance. Images in the database are annotated using the Amazon Mechanical Turk (AMT) platform. The resulting set contains* 18,313 *images with* 23,788 *annotated people. The goal of this paper is to present the EMOTIC database, detailing how it was created and the information available. We expect this dataset can help to open up new horizons on creating systems able of recognizing rich information about people's apparent emotional states.*

1. Introduction

When we look at a person it is very easy for us to put ourselves in her situation, and even to $feel$, to some degree, things that this person appears to be feeling. We use this exceptional ability of guessing how others feel constantly in our daily lives. Such *empathizing* capacity serves us to be more helpful, sensitive, sympathetic, affectionate and cordial in our social interactions. More generally, this capacity help us to understand better other people, to understand the motivations and goals behind their actions and to predict how they will react to different events.

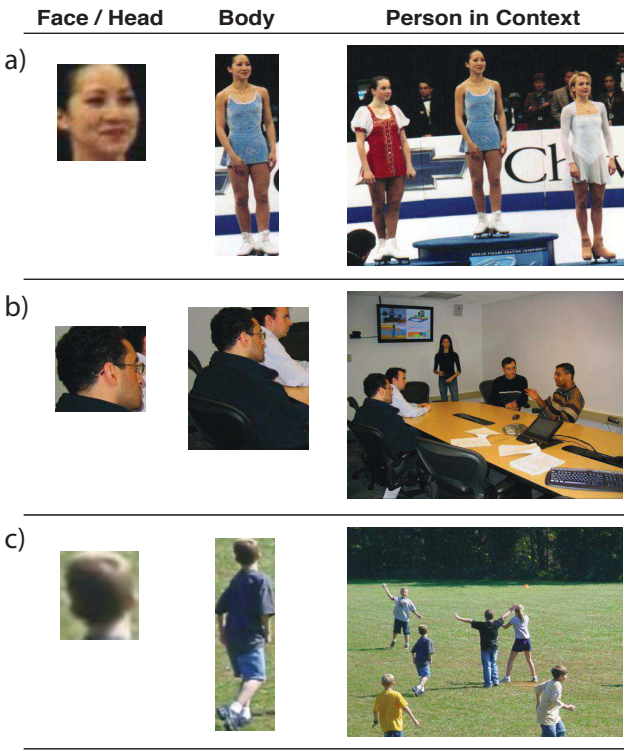

Figure 1: What can we estimate about these people emotional states?

In this paper we introduce the EMOTIC (Emotions in Context) database. The EMOTIC database is a large scale annotated image database with people in context. In this database people are annotated according to their apparent emotions, with a rich set of 26 emotional categories and also with continuous dimensions (valence, arousal and dominance). The images, which show context of the person,

cover a wide range of scene types and activities, allowing the study of emotion recognition beyond the analysis of facial expressions.

There has been a lot of research in emotion recognition from images. In particular, remarkable research has been done in the direction of recognizing the 6 basic emotions [11] (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*), mainly from facial expression, but also from body language analysis. In section 2, we give a brief overview of some of the most relevant works. However, despite these efforts, machines are still far from recognizing emotional states as we do.

The problem of emotional state recognition is extremely complex, but our hypothesis is that there are two main important limitations in the current approaches: (1) First, most of the existing databases in emotion recognition lack of fine-grain labels of human emotions. Most studies classify emotions according to 6 categories, but this is far from the fine grain categorization that humans are capable of. In this work we introduce a more sophisticated set of 26 emotion categories and combine them with the common continuous dimensions (valence, arousal and dominance). This combination provides a rich description of the emotional state of a person. (2) Second, the context (the surroundings of the person) is an important source of information and has been traditionally dismissed. The EMOTIC database attempts to overcome these two issues.

Recent research in psychology highlights the importance of context in the perception of emotions [4]. The importance of the context to infer fine grain information on apparent emotional states is illustrated in Figure 1. For example, in Figure 1.a, when we look only at the face of the person (first column), we can guess that the person is feeling *Happiness*, but is hard to infer additional information about her emotions. The body pose and clothing (second column) gives additional clues and we can infer that she is practicing some sport. However, when we consider the whole context (third column), we can see that she was involved in a competition and she got the first prize. From this information, we can say she probably feels *Excitement* and *Confidence*. A similar story can be told for the person in Figure 1.b. We only see part of the face (first column) which is not very informative, but the body (second column) indicates that the person is looking away toward something or someone - which apparently has his attention. Even now we cannot tell much. It is only when we look at the whole context (third column) that it becomes clear the person is in a meeting room and he is paying attention to a person talking, probably feeling *(Engagement)*. Figure 1.c shows even a more challenging situation. We just see the back of the person's head (first column) which does not give any information about the emotional state of the person. The body pose (second column) reveals part of the story, but it is only the whole image (third column) that tells us that the boy is playing, so he probably feels *Engagement* with some other kids, and he is probably in a state of *Anticipation* to the trajectory of the ball. The goal of the EMOTIC dataset is to provide data for developing automatic systems that can make these type of inferences.

The EMOTIC database is introduced in a paper accepted for the Computer Vision and Pattern Recognition 2017 conference [18]. In this paper we give more details on the dataset, on the image annotation, and on the annotation consistency among different annotators. We present also extended statistics and algorithmic analysis of the data using state-of-the-art methods for scene category and scene attribute recognition. Our analytics show different distribution patterns of emotions depending on the different scenes and attributes. The obtained results indicate that current systems of scene understanding can be used to incorporate the analysis of context in the understanding of people's emotions. Thus, we think that EMOTIC dataset, in combination with previous datasets for emotion estimation (see Section 2.1 for an overview), can help to make further steps in the direction of designing systems capable of recognizing people's emotions as humans do.

2. Related Work

Most of the research in computer vision to recognize emotional states is contextualized in facial expression analysis (e.g.,[5, 13]). 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 [5].

Instead of recognizing emotion categories, some recent works on facial expression [27] 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. [26] used the body pose to recognize the 6 basic emotions, performing experiments on a small dataset of non-spontaneous poses acquired under controlled conditions.

2.1. Related datasets

In recent years we observed a significant emergence of affective datasets to recognize people's emotions. The GENKI database [1] contains frontal face images of single person with wide ranging illumination, geographical, personal and ethnic setting and the images are labeled as *smiling* or *non-smiling*. The ICML face-expression recognition dataset [2] consists of $28k$ images annotated with 6 basic emotions and a neutral category. The UCDSEE dataset [28] has a set of 9 emotion expressions acted by 4 persons acquired using strictly the same lab setting in order to focus mainly on the facial expression of the person.

The dynamic body movement is also an essential source of emotion. The studies [16, 17] establish the relationship between affect and body posture using as ground truth the base-rate of human observers. The data used consists of a spontaneous subset acquired under a controlled setting (people playing Wii games). The GEMEP database [3] is a multi-modal (audio and video) dataset and comprises 10 actors playing 18 affective states. The dataset has videos of actors showing emotions through acting - body pose and facial expression combined.

EMOTIW challenges [7] hosts 3 databases *viz.* 1) *The AFEW* database [6] focuses on emotion recognition from video frames taken from movies and TV shows; where the actions are semi-spontaneous and are annotated with attributes like name, age of actor, age of character, pose, gender, expression of person, the overall clip expression and the basic 6 emotions and a neutral category; 2) *The SFEW* dataset [8] is a subset of AFEW database consisting of images of face-frames annotated specifically with the basic 6 emotions and a neutral category *and*, 3) *The HAPPEI* database [9] focuses on the problem of group level emotion estimation and it is a first attempt to 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 emotion categories for people, such as *happy* and *curious*. These attributes show some overlap with the categories that we define in this paper. However, COCO attributes are not intended to be exhaustive for emotion recognition, and not every person in the dataset is annotated with affect attributes.

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 emotion categories with definitions.

3. EMOTIC Dataset Creation

Our aim was to create a database of natural images, capturing the subjects and their contexts with their natural unconstrained environments. We started by collecting images from well established datasets like MSCOCO [20] and ADE20K [33]. These datasets host a good number of images which satisfy our criteria. We also downloaded images after searching on Google search engine. We used various combination of words representing a varied mixture of subjects, locations, situations and contexts. This resulted in a challenging collection of images, that combine images of people under different situations, performing different tasks and showing a wide range of emotional states. Currently, the EMOTIC database consists of 18316 images with 23788 people annotated. EMOTIC dataset is split in training (70%), validation (10%), and testing (20%) sets.

3.1. Emotion Representation

EMOTIC dataset combines 2 methods to represent emotions:

- **Discrete Categories**: We define an extended list of 26 emotional categories. Table 1 gives the definition of each emotion category. Two examples of images for each of the emotion category are shown in Figure 2.
- **Continuous Dimensions**: We also used the *VAD* Emotional State Model to represent emotions. 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 Figure 4 we show images of the EMOTIC database along with their annotations.

To define the 26 emotion categories, we collected a vocabulary of affective states. Using word connections (synonyms, affiliations, relevance) and the inter-dependence of a group of words (psychological research and affective computing [14, 25]), we started forming word-groupings. After multiple iterations and cross-referencing with dictionaries and research in affective computing, we obtained the final 26 categories (ref Table 1). While forming this group of 26 emotion categories, we adjudged it necessary for the group to follow two important conditions: *(1) Disjointness* and *(2) Visual Separability*. By *disjointness*, we mean that 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). By *visual separability* we mean that two affective states were assigned to the same emotion group in case we find, qualitatively, that the two could not be visually separable under the conditions of our database. For example, in Figure 2 the images for *7.Happiness* and *1.Peace* show clearly the visual separability of these categories in spite of being similar. However, the category *excitement*, for instance, 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 26 categories also include the 6 basic emotions (categories 7, 10, 18, 20, 22, 24) defined by Ekman [11]. Note that *Aversion* is a general form of the basic emotion *Disgust*, hence it makes more sense to keep it as a main emotion category.

3.2. Image Annotation

Images have been annotated on the Amazon Mechanical Turk(AMT) Platform. Figure 6 shows the two different annotation interfaces created for labeling images with emotion categories (Figure 6.a) and continuous dimensions (Figure 6.b). In the AMT interface for continuous dimensions we also asked AMT workers to annotated the gender and the estimated age range of the person in the bounding box according to the following age ranges: kid (0-12 years old), teenager (13-20 years old), adult (more than 20 years old).

We adopted two methods to control the annotation quality, apart of providing the AMT workers with extended annotation instructions and examples. First, we launched a qualification task to monitor the annotation of one image according to discrete categories and one image according to continuous dimensions. In each of the control figures we manually selected all the categories that we thought were not clearly apply for that image, and also those ranges of continuous dimensions that were, in our opinion, out of an acceptable response. For instance, in the image shown in 6, a worker that selected $pleasure$, $disconnection$, $sadness$, $fear$, $pain$, $suffering$, or a worker that selects $1 - 2$ for valence, arousal or dominance would not pass. We considered that a worker that labeled according to these light restrictions understood well the task and did not make a random labeling. Just those workers that label the control images under the the mentioned restrictions were allowed to label images of the dataset. Secondly, we insert 2 control images, with restrictions similar to the ones mentioned before, for every 18 images to track the consistency during the annotation. The annotations of our database come from those workers that keep passing the control images during their HITs.

3.3. Annotation Consistency

The Validation set was annotated by 5 different workers, to check the annotation consistency among different people. Although there is an inherent subjective nature in this annotation task, we observed that there is a degree of agreement among different annotators. To measure this agreement quantitatively, we computed *Fleiss' Kappa* (κ) statistic using the 5 annotations of each image in the validation set. The obtained result is $\kappa = 0.31$. Furthermore, more than 50% of the images in the validation set has $\kappa > 0.31$, which indicates a much higher agreement level than random chance (notice that random annotations will produce $\kappa \sim 0$).

For every image in validation set, there is at least one category which is annotated by 2 or more annotators. In 60% images (validation), there is at least one category which is annotated by at least 3 annotators. These statistics which are reflected in the Figure 5 are a good indicator of the agreeableness amongst the human annotators.

For continuous dimensions, the standard deviation among the 3 different workers is, in average, 1.41, 0.70 and 2.12 for valence, arousal and dominance respectively. Dominance shows a higher dispersion in its value as compared to

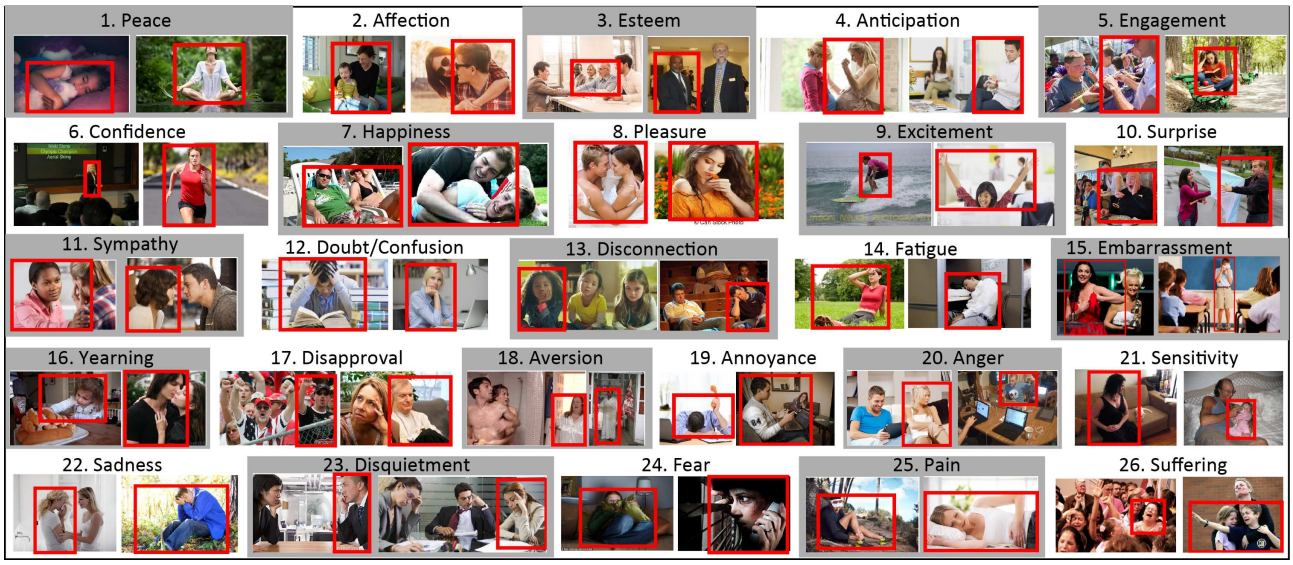

Figure 2: Visual examples of the 26 emotion categories defined in Table 1. Per each category we show two images where the person marked with the red bounding box has been annotated with the corresponding category.

Figure 3: Examples of images from the EMOTIC dataset with different scores of Valence, Arousal, and Dominance.

other dimensions. This means that the agreement deviation is higher for dominance than for other dimensions. Note that the range of values for each dimension is between 1 to 10 - 1 being the lowest and 10 the highest level.

4. Dataset Statistics

Of the 23,788 persons annotated in the dataset, 66% are males and 34% are females. The age distribution of the annotated people is the following: 11% children, 11% teenagers, and 78% adults.

Figure 7.a shows the number of people for each emotional category, while Figures 7.b, 7.c and 7.d show the number of people for valence, arousal and dominance continuous dimensions for each score, respectively.

It is interesting to see how the values (1 - lowest to 10 - highest) of a given continuous dimension are spread across the dataset for each emotional category. Figure 8.a shows

Figure 4: Annotated images from the EMOTIC dataset.

the plot for the spread of *Valence* values across each one of them. The categories are sorted according to the mean value of *Valence* across each category. We can clearly see that the lowest mean of *Valence* is for emotions like *Suffering, Pain, Sadness* - indicating that *Valence* values, in average, are low for these categories. This shows consistency on the annotations of our EMOTIC dataset, since one would expect that images with emotion categories like *Suffering, Pain, Sadness* should exhibit, in average, low levels of *Valence*. Similarly, it is clear that emotions like *Happiness,*

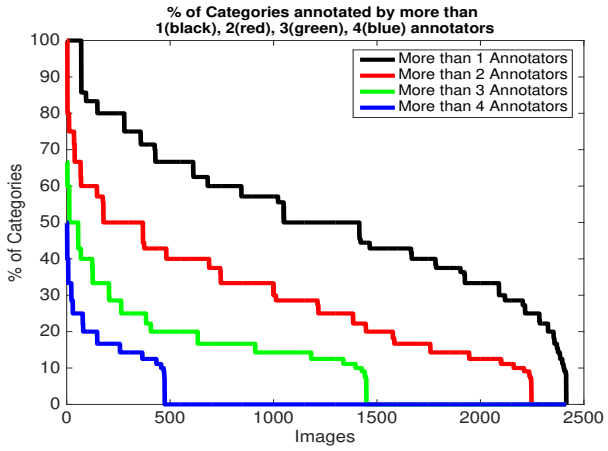

Figure 5: Images of validation set with at least one of all the annotated categories with more than (1,2,3,4) annotators

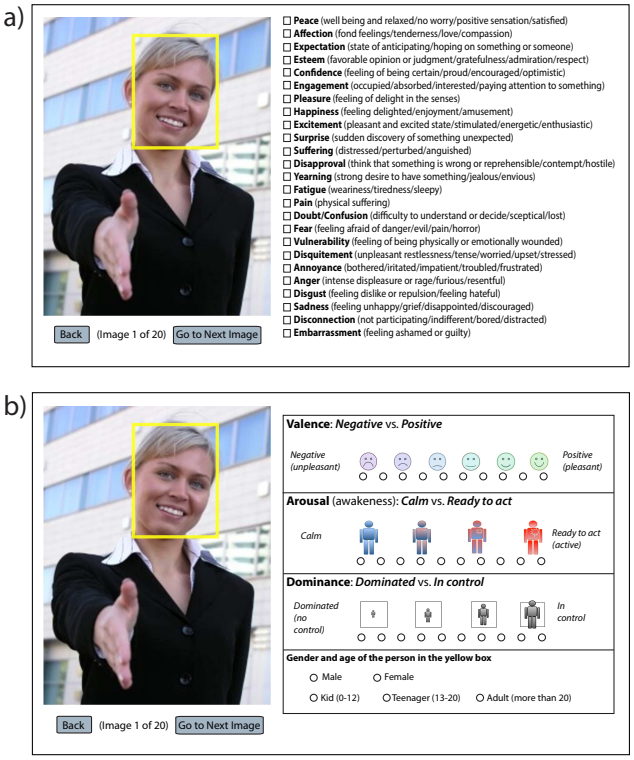

Figure 6: AMT Interfaces for emotion category (a) and continuous dimensions (a) annotation.

Affection, Pleasure should have a higher level of *Valence* and this fact is apparent from the plot itself. Finally, as expected, the plot also shows the correlation between a person in a disconnected emotional state and a mid-level *Valence* value - *Disconnection* lies in the mid-range of the sorted emotion categories. Similarly, in Figure 8.b we can see the same type of information for the *arousal* dimension. Notice

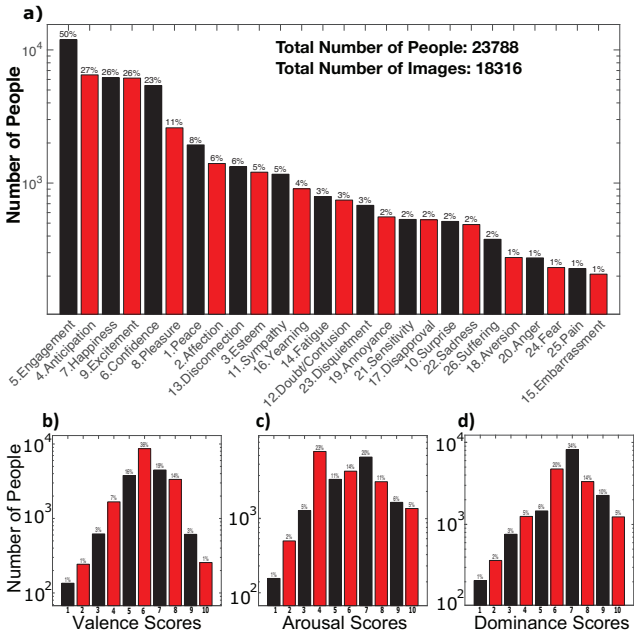

Figure 7: a) Number of people per emotion category in the EMOTIC dataset; b),c),d) Number of people per score in each of the continuous dimensions.

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

We also computed the co-occurrence probability of categories. These probabilities are shown in Figure 9. Given a row, corresponding to the emotional category c_1, and a column, corresponding to the emotional category c_2, each entry corresponds to the probability $P(c_2|c_1)$. From these results we can observe interesting patterns of category co-occurrences. For instance, we 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 category annotations and we ob-

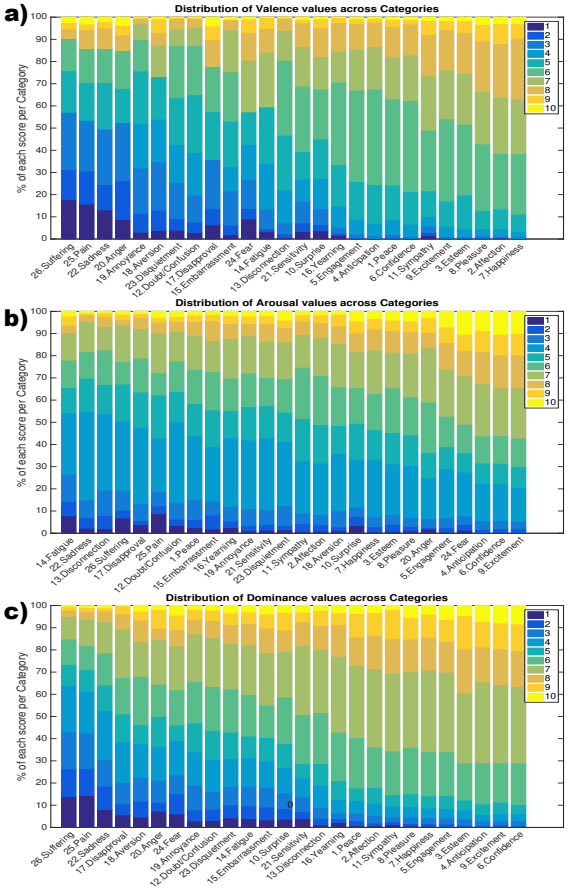

Figure 8: Per each of the continuous dimensions, distribution of the scores across the different emotion categories.

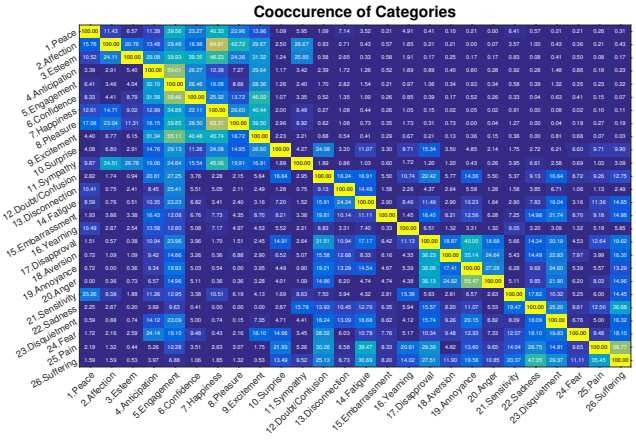

Figure 9: Category co-ocurrance probabilities.

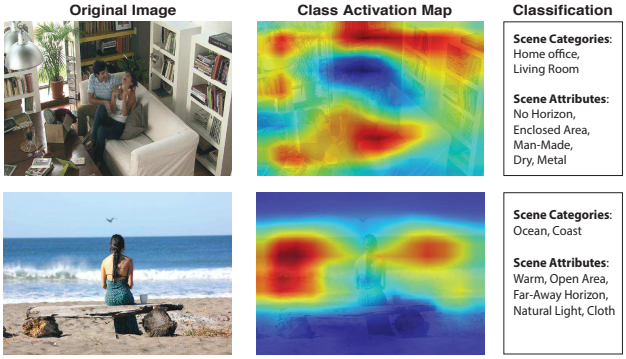

Figure 10: Examples of scene category and attribute recognition in images from the EMOTIC dataset, with the corresponding class activation map.

served that some category groups appear frequently in the EMOTIC database. Some examples are {*anticipation*, *engagement*, *confidence*}, {*affection*, *happiness*, *pleasure*}, {*doubt/confusion*, *disapproval*, *annoyance*}, {*yearning*, *annoyance*, *disquietment*}.

5. Dataset Algorithmic Analysis

We classified the scenes of EMOTIC dataset using a "state-of-the-art Convolutional Neural Network model for scene recognition [32]. Figure 10 shows two examples of scene category and scene attribute recognition in images of EMOTIC dataset. The figure shows the original images, the class activation maps [30] (that correspond to the region of the image that supports the decision of the classifier, and the scene categories (top 2) and scene attribute (top 5) automatically recognized in each image. As we can see, the results are very accurate. In general, as reported by the authors in [31], the recognition accuracy of these systems is around 78%, according to the feedback provided by the users of online demo for scene recognition.

Using the classification labels obtained with these CNN models we can now study patterns on emotion distribution in different scene categories and under different scene attributes. Using our data, we computed the probability of each emotion at each place as $P(emo|place) = Nemo/Npeople$, where $Npeople$ is the number of people that we observed in the specific place category, and $Nemo$ is the number of those people that have been labeled with the specific emo category. Figure 11 shows representative examples of the emotion category distribution in different scenes. We see that, in our data, there are different patterns. For instance, the most frequent emotions in a bedroom are $engagement$, $happiness$, $pleasure$, $peace$, and $affection$, while in a baseball field, the most frequent emotions are $engament$, $anticipation$, $confidence$, and $excitement$. Among these examples, we also see that emotions like $disquietment$, $annoyance$, or $anger$ are more frequent in bedrooms (where people have some privacy degree), in offices (where people can feel tired of working) or

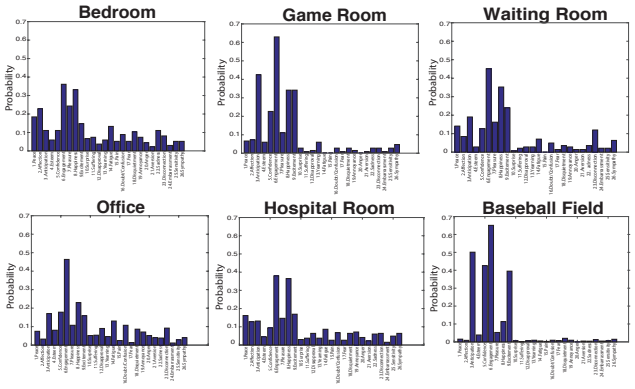

Figure 11: Emotion Distribution per place category

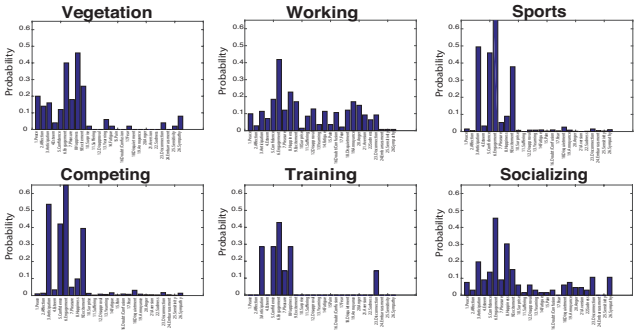

Figure 12: Emotion Distribution per place attribute

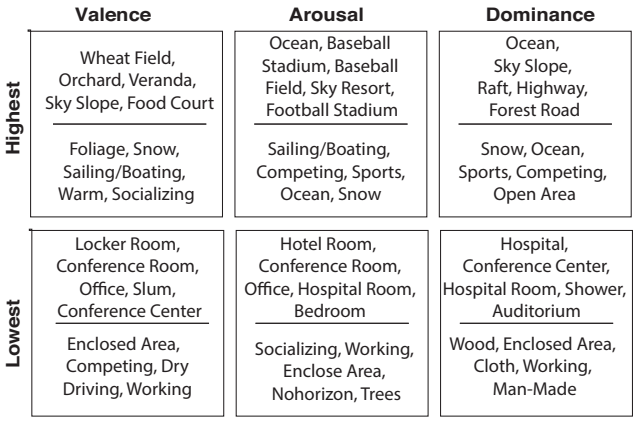

Figure 13: Place categories and attributes with highest and lowest average scores for valence, arousal, and dominance.

in hospital rooms (where people can be worried).

We can do a similar analysis using the recognized scene attributes. Representative results are shown in Figure 12. Here, we also observe some interesting patterns. Among these examples, we see that the category $peace$ is significantly higher when scenes have the attribute $vegetation$. This seems reasonable, since natural areas are more suitable for relaxing. Similarly, the highest frequency for $sympathy$ is shown for the attribute $socializing$, followed by the attribute $vegetation$. Taking into account the 6 attributes shown in Figure 12, it seems reasonable to see more frequently $sympathy$ in scenes with these two attributes than, for instance, in scenes of $sports$ or $competing$. We also observe very high frequencies for the categories $engegment$, $anticipation$, $confidence$, and $excitement$ in scenes with the attributes $sports$ and $competing$. Notice that, in general, the most frequent emotions correlate with the most frequent emotions in the whole database (see Figure 7), as expected. For instance, there are many people in the database showing $engagement$, since usually people are "involved into something." As a consequence, we see a high frequency of the emotion $engagement$ in all of the scene categories and attributes. However, even though the main pattern of the general emotion distribution is preserved, we see specific variations in each scene category and attribute.

Regarding the continuous dimensions, we also observe interpretable patterns. Figure 13 shows an overview of scene categories (top section of each box) and scene attributes (low section of each box) with highest and lowest valence, arousal, and dominance, respectively, in average.

6. Conclusions

In this paper we present the EMOTIC database, a database of images with people in context annotated according to their apparent emotional states. The EMOTIC database combines 2 different annotation approaches: an extended set of 26 emotional categories and a set of 3 continuous dimensions (Valence, Arousal, and Dominance). In this paper we present the details on the EMOTIC database creation and statistics. We also present an algorithmic analysis of the data performed using state-of-the-art methods for scene category and scene attribute recognition. Our results suggest that current scene understanding techniques can be used to incorporate the analysis of the context for emotional states recognition. Thus, the EMOTIC dataset, in combination with previous datasets on emotion estimation, can open the door to new approaches for apparent emotion estimation in the wild from visual information.

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] GENKI database. `http://mplab.ucsd.edu/wordpress/?page_id=398`. Accessed: 2017-04-12.

[2] ICML face expression recognition dataset. https://goo.gl/nn9w4R. Accessed: 2017-04-12.

[3] T. Bänziger, H. Pirker, and K. Scherer. Gemep-geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions. In *Proceedings of LREC*, volume 6, pages 15–019, 2006.

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

[5] 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.

[6] A. Dhall et al. Collecting large, richly annotated facial-expression databases from movies. 2012.

[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*, pages 427–432. ACM, 2016.

[8] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2106–2112. IEEE, 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] 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.

[17] 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.

[18] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotion recognition in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[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. Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[26] 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.

[27] 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.

[28] J. L. Tracy, R. W. Robins, and R. A. Schriber. Development of a facs-verified set of basic and self-conscious emotion expressions. *Emotion*, 9(4):554, 2009.

[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. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[31] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016.

[32] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.

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