Automatic RGB Inference Based on Facial Emotion Recognition

Nicolo' Brandizzi[1], Valerio Bianco[1], Giulia Castro[1], Samuele Russo[2] and Agata Wajda[3]

[1]*Department of Computer, Control and Management Engineering, Sapienza University of Rome, Via Ariosto 25, 00135, Rome, Italy*
[2]*Department of Psychology, Sapienza University of Rome, Via Ariosto 25, 00135, Rome, Italy*
[3]*Department of Mathematics Applications and Methods for Artificial Intelligence, Faculty of Applied Mathematics, Silesian University of Technology, 44-100 Gliwice, Poland*

Abstract

Recently, Facial Emotion Recognition (FER) has been one of the most promising and growing field in computer vision and human-robot interaction. In this work, a deep learning neural network is introduced to address the problem of facial emotion recognition. In particular, a CNN+RNN architecture has been designed to capture both spatial features and temporal dynamics of facial expressions. Experiments are performed on CK+ dataset. Furthermore, we present a possible application of the proposed Facial Emotion Recognition system in human-robot interaction. A method for dynamically changing ambient light or LED colors, based on recognized emotions is presented. Indeed, it is proven that equipping robots with the ability of perceiving emotions and accordingly reacting by introducing suitable emphatic strategies significantly improves human-robot interaction performances. Possible scenarios of application are education, healthcare and autism therapy where such kind of emphatic strategies play a fundamental role.

Keywords

Facial Emotion Recognition, Human Robot interaction,

1. Introduction

Facial Emotion Recognition (FER) is the process of identifying human feelings and emotions from facial expressions. Nowadays, automatic emotion recognition has a key role in a wide area of applications, with particular interest in cognitive science[1], human-robot and human-computer interaction. In human-robot interaction for example, the ability of recognize intentions and emotions and accordingly react to the particular motivational states of the user is crucial for making interaction more friendly and natural, improving both usability and acceptability of the new technology.

Ekman in [2] defined a set of six universal emotions: anger, disgust, fear, happiness, sadness, and surprise, which can be universally recognized and described regardless of people culture and context by using a set of action unit (AU) described in table 1. Each AU is the action of a muscle of the face that is typically activated when a given facial expression is produced. For example, AU number 1, that corresponds to "Inner brow raiser" typically appears when people show surprise emotional state, together with jaw drop. Facial emotion recognition is a challenging task due to interclass similarities problems. Indeed different people can show emotions in a different and personal way and with a different level of intensity which makes the problem particularly hard. On the other hand it is possible that different motivational states show very similar features and similar facial expressions.

In this work, a standard CNN-RNN architecture is used to learn both spatial and temporal cues from human facial expressions built up gradually across time. A method to express emotions by dynamically changing RGB ambient light components based on recognized emotional state is here proposed. Indeed, a lot of studies have been conducted in order to understand the relationship between colors and emotions, and precisely how a particular color can evoke positive feelings in the observer. So depending on the particular recognized human emotion a different ambient light color will be set according to Nijidam color-emotion mapping theory [3].

The remainder of this paper is structured as follows. Section 2 analyzes existing literature and discusses the state-of-the-art in facial emotion recognition. Section 3 is dedicated to the description of the dataset and data refinement process. Section 4 formalizes the problem and explains proposed CNN+RNN model architecture. Section 5 shows experiments, implementation details, and achieved results. Section 6 propose a real-time human-robot interaction application which can take advantage and multiple benefits from the proposed Facial Emotion Recognition system. Finally, in Section 7, we discuss

SYSTEM 2021 @ Scholar's Yearly Symposium of Technology, Engineering and Mathematics. July 27–29, 2021, Catania, IT

brandizzi@diag.uniroma1.it (N. Brandizzi); castro.1742813@studenti.uniroma1.it (G. Castro); samuele.russo@studenti.uniroma1.it (S. Russo); agata.wajda@polsl.pl (A. Wajda)
0000-0002-1846-9996 (S. Russo); 0000-0002-1667-3328 (A. Wajda)

Table 1
Facial Action Unit (AU) code and corresponding description.

AU	Description	AU	Description	AU	Description
1	Inner Brow Raiser	13	Cheek Puller	25	Lips Part
2	Outer Brow Raiser	14	Dimpler	26	Jaw Drop
4	Brow Lowerer	15	Lip Corner Depressor	27	Mouth Stretch
5	Upper Lip Raiser	16	Lower Lip Depressor	28	Lip Suck
6	Cheek Raiser	17	Chin Raiser	29	Jaw Thrust
7	Lip Tightener	18	Lip Puckerer	31	Jaw Clencher
9	Nose Wrinkler	20	Lip Stretcher	34	Cheek Puff
10	Upper Lip Raiser	21	Neck Tightener	38	Nostril Dilator
11	Nasolabial Deepener	23	Lip Tightener	39	Nostril Compressor
12	Lip Corner Puller	24	Lip Pressor	43	Eyes Closed

conclusions and future works.

2. Related Work

Several techniques and deep learning models have been investigated over the last decade in order to address the problem of facial analysis and emotion recognition from RGB images and videos. Most of them use Convolutional Neural Networks (CNNs) for extracting geometric features from facial landmark points. In order to train such high-capacity classifier, with the very small-size available FER dataset, one of the most common approach is to use transfer learning. Works [4, 5] use pre-trained models such as VGG16 and AlexNet to initialize the weights of the CNN that can improve accuracy and reduce overfitting. Nguyen *et al.* in [4] introduced a two-stage supervised fine-tuning: a first-stage fine-tuning is applied using auxiliary face expression datasets followed by a final fine-tuning on the target AFEW dataset [6]. In [5] a VGG-16 deep pre-trained model plus redefined dense layers is used in FER, by identifying essential and optional convolutional blocks in the fine-tuning step. In the training process the selected blocks of VGG-16 model are included step by step, instead of training all at a time, to diminish the effect of initial random weight.

Instead of analyzing static images independently, thus ignoring the temporal relations of sequence frames in videos, also 3D CNNs were explored in order to extract spatio-temporal features with outstanding results. In [7], 3D CNNs was used to model appearance and motion of videos, learning simultaneously spatial and temporal aspects from image sequences. In [8], two deep networks are combined: a 3D CNN is used to capture temporal appearance of facial expressions, while a deep temporal geometry network extracts geometrical behaviours of the facial landmark points.

Similarly, some recent works have proposed to use both combination of CNNs and RNNs capable of keeping track of arbitrary long-term dependencies in input sequences. In [9] multiple LSTMs layers are stacked on top of CNNs. Then temporal and spatial representations are aggregated into a fusion network to produce per-frame prediction of 12 facial action units (AU). Fan *et al.* in [10] propose an hybrid network combining a CNN-features-based spatio-temporal RNN model with a 3 dimensional Convolutional Neural Network (C3D), including also audio features in order to maximize accuracy predictions. To deal with expression-variations and intra-class variations, namely intensity and subject identity variations, [11] introduces objective functions on CNN to improve expression class separability of the spatial feature representation and minimize intra-class variation within the same expression class. Differently from previous works, which first apply CNN architectures or pre-trained image classifier as visual feature extractor, and then use extracted spatial feature representation for training the RNNs separately, the proposed network want to analyze the spatio-temporal behaviour of facial emotions by using an end-to-end trainable CNN+RNN computational efficient architecture.

Several experiments have also been done to prove how a facial emotion recognition systems can improve human-robot interaction performances and potentialities. Jimenez *et al.* in [12] show how monitoring colored lights based on user emotion can represent a real communication channel between humans and robots. They introduce a self-sufficiency model system that recognizes and empathizes with human emotions using colored lights on a robot's face. Feldmaier *et al.* in [13] also show the effectiveness of displaying color combinations and color patterns in Affective Agents, also adjusting variations of intensity, brightness and frequency to obtain a psychological influence in the user. A similar strategy is adopted in this work as well.

Many other works have been recently published in the field of emotion recognition, face detection, and related classification tasks[14, 15].

Table 2
Ekman's six basic emotions description in terms of facial Action Units. With reference to table 1 we define 1: Inner Brow Raiser, 2: Outer Brow Raiser, 4: Brow Lowerer, 7: Lip Tightener, 9: Nose Wrinkler, 10: Upper Lip Raiser, 12: Lip Corner Puller, 15: Lip Corner Depressor, 17: Chin Raiser, 20: Lip Stretcher, 24: Lip Pressor, 25: Lips Part, 26: Jaw Drop.

Emotional state	Action Units
Anger	4, 7, 24
Disgust	9, 10, 17
Fear	1, 4, 20, 25
Happiness	12, 25
Sadness	4, 15
Surprise	1, 2, 25, 26

Table 3
Final number of samples per category in the dataset. In order: Anger (Ang), Neutral (Neu), Disgust (Dis), Fear (Fea), Happiness (Hap), Sadness (Sad), Surprise (Sur) number of videos samples.

Emotion label	Ang	Neu	Dis	Fea	Hap	Sad	Sur
#Videos	51	52	70	61	108	82	87

3. Dataset

For training, validating and testing the Facial Emotion Recognition model the Extended Cohn-Kanade Dataset (CK+) [16] was used. It contains 593 sequences across 123 subjects and each of the sequences contains images from onset (neutral frame) to peak expression (last frame). The image sequence can vary in duration from a minimum of 10 to a maximum of 60 frames. Images have frontal views and 30-degree views and were digitized into either 640x490 or 640x480 pixel arrays with 8- bit gray-scale or 24-bit color values. Each of the image sequences is labelled with Action Unit combinations. A complete list of the possible AUs is reported in table 1.

For each of the data sequence, if the action units list show consistency with one of the six basic emotion category among Anger, Disgust, Fear, Happiness, Sadness, Surprise and Contempt, a nominal emotion label is associated to the sequences. At this aim, table 2 shows a complete mapping between Ekman's six basic emotions and AUs. Note that only the six basic emotions are considered in the proposed FER model and then reported in table 2, while *Contempt* category which is very similar to *Disgust* emotion class and do not belongs to basic emotions group is excluded. As a result of this selection process, only 296 of the 593 sequences fit the prototypic definition and meet criteria for one of the six discrete emotions. Prototypes definitions used for translating AU sores into emotions terms are shown in Table 6 in the Appendix. Comparison with the emotion prediction Table 6 was done by applying the emotion prediction rule very strictly.

In order to maximize the amount of data which appears to be too poor for the training model, another 35% of the dataset was hand-made labelled by using not only Prototypes but also their Major Variants as shown in the Emotion Prediction Table 6. Compared with Prototypes, variants allow for subset of AUs. As a consequence they are less strictly definitions but always truly representative for the given emotion. As a result, a total of 511 video sequences are collected by including also the major variants definitions in the conversion rules. Also neutral facial expressions were included, for a total of 7 emotions categories: Neutral (Neu), Anger (Ang), Disgust (Dis), Fear (Fea), Happiness (Hap), Sadness (Sad) and Surprise (Sur). Notice that even if the dataset increases its dimension, it remains unbalanced, indeed some of the categories such as surprise and happiness are more represented in the dataset if compared to the others. The final distribution of data samples among the seven categories is reported in table 3. The minimum length of data samples in the dataset is 10 frames. For this reason, in order to maximize the number of input sequences, and collect the largest amount of training data, the sequence length, i.e the number of frames per video is set to 10. For sequence of grater length, only the last 10 frames are considered, in order to ensure that the apex (peak) frame that is the most representative for the given emotion will be captured. For each of the frames pixels values are rescaled in order to have each pixel $\in [0, 1]$. A central cropping is applied to each of the frame in order to have a more focus on human face. After resizing and cropping the final dimension of each input sequences will be

$$(n\_frames, width, height, n\_channels) = (10, 48, 48, 3)$$

4. Model Architecture

The emotion recognition task is modelled as a multi-class classification problem over a set of 7 different categories $Y = \{\text{Anger, Neutral, Disgust, Fear, Happiness, Sadness, Surprise}\}$.

As shown in sec 2 Recurrent Neural Networks (RNN) in combination with Convolutional Neural Networks (CNN) and 3 dimensional Convolutional Neural Network (C3D) provide powerful results when dealing with sequential image data. Following this approach, a standard CNN + RNN computational efficient architecture is here proposed. From the input sequence of frames, spatial feature representations are learned by a Convolutional Neural Network (CNN). In order to capture the facial expression dynamics, temporal features representation of the

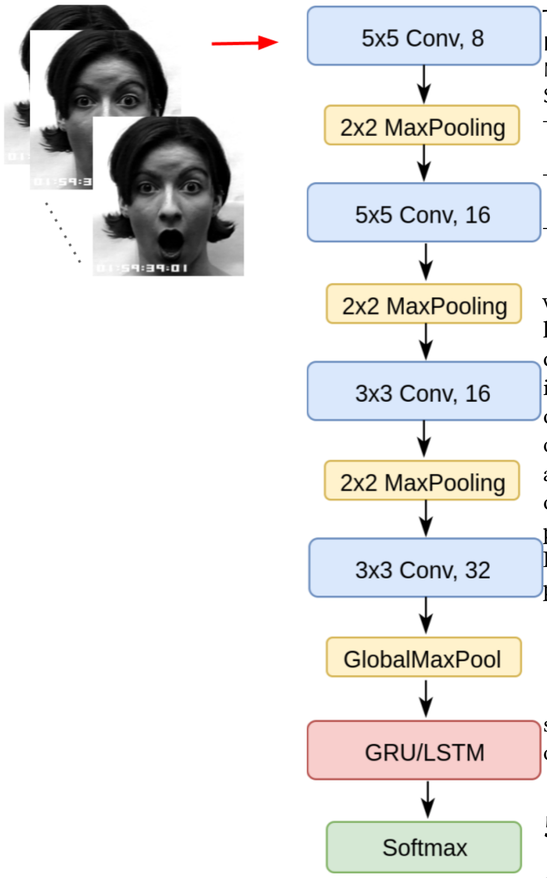


Figure 1: Proposed model Architecture for Facial Emotion Recognition system which takes as input a 10 frames sequence and outputs a single label emotion.

facial expression is learned via the Recurrent Neural Network (RNN). Both Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been tested for the given problem and provide comparable results. However GRU have a less complex structure and are computationally more efficient. Therefore given also the fact that the model does not have a huge amount of data, GRU are preferred to get a good accuracy. Moreover, such 2D CNN + GRU approach allows end-to-end trainable model which is lower computational expensive comparing with the others. Indeed the whole model counts a very small number of parameters, around 8k, which makes it very light.

The model takes as input a window of 10 RGB frames of size 48x48. Data are processed in batches of 24 image sequences. The full architecture of the proposed model is shown in Figure 1. It consist of a four-layer 2D CNN. The first two convolutional layers have 5x5 kernel size, while last couple have 3x3 filter size. All 2D Max Pooling have a kernel size of 2x2. In order to extract temporal correlation in the extracted input features, CNN output is directly fed as input to a GRU of 8 units, which is the output dimension. Finally, the output layer is a Dense one, that is a deeply fully connected layer, with a *Softmax* activation function. Softmax layer have the same number of nodes as the output layer in order to assign decimal probabilities with sum 1 to each of the emotion categories. For each value z_i from the neurons of the output layer, per category probability is computed as:

Table 4
Per Category Validation Accuracy. In order: Anger (Ang), Neutral (Neu), Disgust (Dis), Fear (Fea), Happiness (Hap), Sadness (Sad), Surprise (Sur) accuracy scores.

Emotion label	Ang	Neu	Dis	Fea	Hap	Sad	Sur
Val Accuracy	10%	83%	50%	30%	86%	86%	79%

$$softmax(z_i) = \frac{exp(z_i)}{\sum_j exp(z_j)} \tag{1}$$

such that probabilities values always sum to 1 and only one emotion label is activated.

5. Experiments

5.1. Implementation Details

To verify the effectiveness of the proposed model, experiments have been conducted on the modified CK+ dataset as described in Section 3.

For an efficient data parsing, input images sequences of 10 frames are first transformed in *TFRecord* files. Then tensorflow *TFRecordDataset* class is used to standardize data and generate batch of 24 images sequences to be fed as input to the model. 80% of the data samples are used in training phase, while the remaining 20% as test set.

For training the model we used *Adam* optimizer with learning rate $1 \times e^{-3}$. For a multi-class classification problem, each of the input sample can only belong to one out of many possible categories, therefore a *Categorical Cross Entropy* Loss was used. Since the dataset is unbalanced the Categorical Accuracy is not enough to have a true evaluation for the model, but also Precision and Recall metrics were considered. In order to prevent over-fitting problem, L1 and L2 Regularizer are set to 0.01 and dropout equal to 0.4. The model was trained for a total of 400 epochs, with a mean training time for a single epoch of only 4sec.

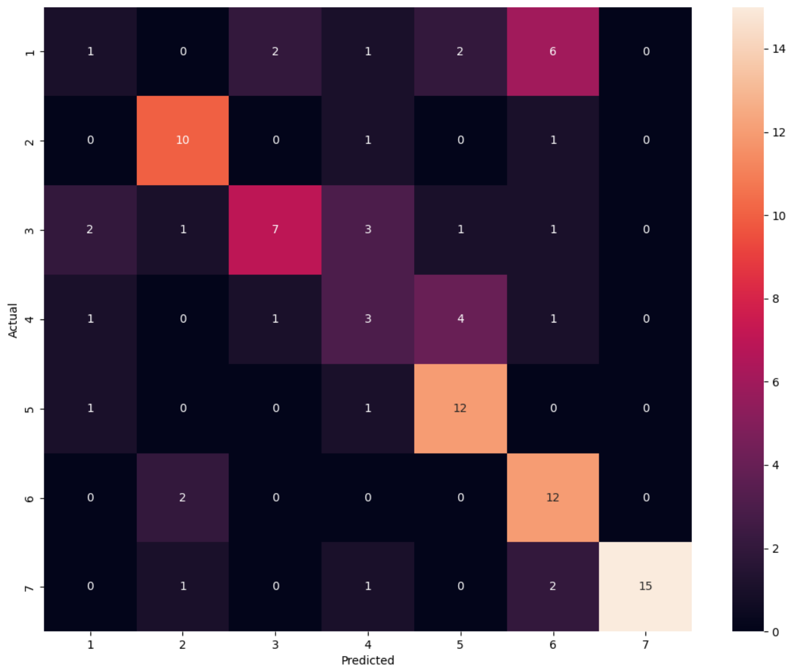


Figure 2: Confusion matrix over CK+ validation set. On the axis motions categories are disposed in the following way: {1:Anger 2:Neutral 3:Disgust 4:Fear 5:Happy 6:Sad 7:Surprise}

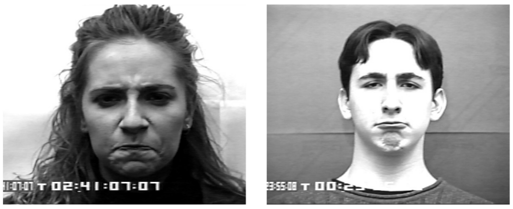


Figure 3: Left: Apex (peak) frame from Angry facial expression from CK+ dataset. Right: Apex frame from Sad facial expression performed by a different subject in CK+ dataset.

Table 5
Evaluation metrics for the model. Mean Validation Accuracy , Precision and Recall across categories.

Validation Accuracy	66%
Validation Precision	66%
Validation Recall	65%

5.2. Results analysis

In this section, the experimental results are presented. Since the dataset is not balanced both precision and recall, together with accuracy score are considered in order to evaluate the performance of the model. For the best validation epoch the model achieve 66% of mean Precision and 65% of mean Recall across categories. Table 4 shows per category validation accuracy rates. As we can notice mean validation accuracy over categories is not so high, indeed per category validation accuracy reaches very high scores for some of the emotion classes and very low values for some others, meaning that the model almost fails when tries to recognize anger and fear emotions, while it behaves very well when dealing with all the others categories. In particular, the model appears very powerful in recognizing Happy and Sad facial expressions with the 86% of accuracy. High accuracy scores are also obtained for Neutral (83%) and Surprise (79%) emotions categories. As a result, even if the model shows very high performances for most of the emotions, difficulties in learning some specific facial expressions, in particular Angry with only the 10% of accuracy, lead to a lower mean accuracy score across categories. To better understand the model, the confusion matrix over the CK+ validation set is reported in Fig 2. As we can notice, the main diagonal is highlighted by high rates of correctly classified samples. However, some off-diagonal elements reflects mislabeled predictions by the classifier.

In particular, predictions errors mainly concern Anger and Fear emotions categories, which appear to be harder to learn and recognize. As shown in the confusion matrix,

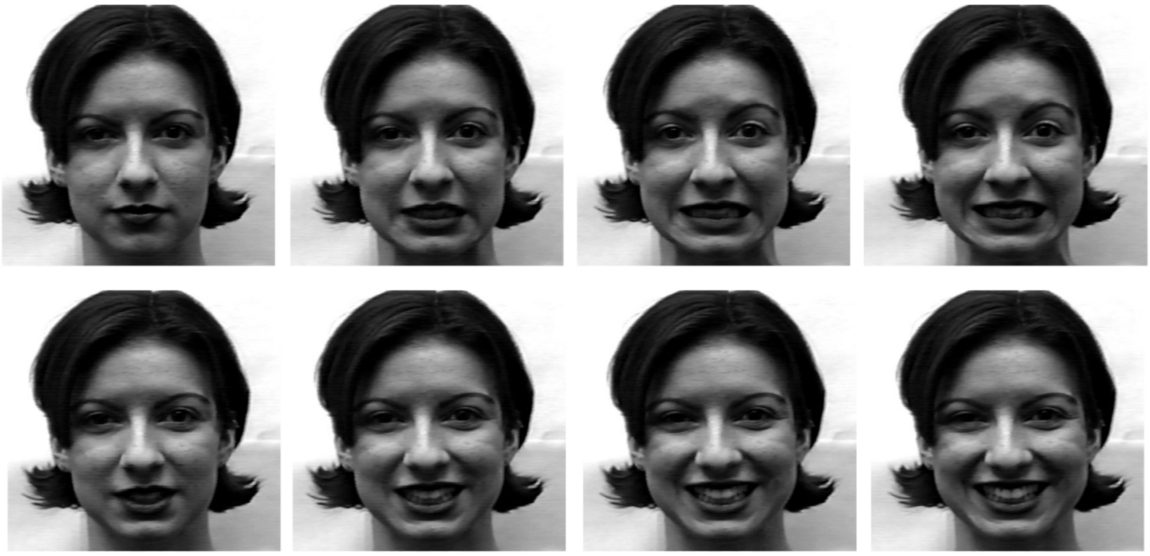


Figure 4: Top row: Fear emotion data sample from CK+ dataset. 4/10 frames are shown to illustrate temporal evolution of the facial expression. Bottom row: Happines emotion evolution over time, performed by the same subject.

Angry facial expressions are frequently confused with Sad ones, which actually looks like very similar. Indeed if we look at some data examples, as shown in figure 3, it will be quite difficult even for humans to infer the correct labeling. This because both Sadness and Anger emotions are characterized by very similar features such as lowered eyebrows and tight mouth as also highlighted in table 2 which shows emotional states description in terms of facial Actions Units. As a consequence, even if Sad facial expressions typically shows much more lowered lip corners, more discrete and shy performed emotions may be easily confused due to inter-class similarity and emotions intensity variations problems. In an analogue way, Fear emotion are often wrongly classified as Happiness, since they show a very similar behaviour of the lips and share action unit number 25 - Lips Part, as reported in table 2. As shown in fig 4 top row, fear facial expression is represented with the jaw dropped open and lips stretched horizontally. As a result, if we look at fig 4, for both Fear and Happiness facial expression (top and bottom rows respectively) lips are closed on the onset (first frame) than start to became farther until to be apart in the apex frame, showing a very similar temporal behaviour which makes it difficult to distinguish between the two.

In summary, the network can precisely recognize Happiness, Sadness, Surprise and Neutral emotional states, while is not so much reliable for Anger and Fear. Therefore only emotions categories with a valid accuracy over the 50% will be used in human-robot interaction applications.

6. Facial Emotion Recognition in human-robot interaction

In this chapter, a possible application of the proposed Facial emotion recognition system in Human-Robot and Human-Computer interaction is presented. More research [17, 18, 19, 20] have proven as equip the robot with the ability of perceiving user feelings and emotions and accordingly react with them can significantly improve the quality of the interaction, and more importantly user acceptability. Affective communication for social robots has been deeply investigated in the last years showing as behind social intelligence, also emotional intelligence plays a crucial role for successfully interactions. In particular the need of recognize emotions and properly introduce emphatic strategies turned out to be fundamental in vulnerable scenarios such us education, healthcare, autism therapy and driving support.

6.1. Application

A very effective and powerful solution in human-robot interaction can be to introduce a colored lights based emphatic strategy. Indeed, colors and emotions are closely linked. Several physical and psychological studies have shown as play with colors and dynamic lights can have very effective outcome since they can evoke feelings and emotions in human observers. For this reason, a method for dynamically changing ambient light or LED colors based on recognized emotions is presented. The aim is to

Table 6

Emotion prediction. Conversion Rules for translating AU scores into Emotions. [16]
Table note: * means in this combination the AU may be at any level of intensity.

Emotion	Prototypes	Major Variants
Surprise	1+2+5B+26	1+2+5B
	1+2+5B+27	1+2+26
		1+2+27
		5B+26
		5B+27
Fear	1+2+4+5*+20*+25	1+2+4+5*+L or R20*+25, 26, or 27
	1+2+4+5*+25	1+2+4+5*
		+2+5Z, with or without 25, 26, 27
		5*+20* with or without 25, 26, 27
Happiness	6+12*	
	12C/D	
Sadness	1+4+11+15B with or without 54+64	1+4+11 with or without 54+64
	1+4+15* with or without 54+64	1+4+15B with or without 54+64
	6+15* with or without 54+64	1+4+15B+17 with or without 54+64
		11+17
		25 or 26 may occur with all prototypes or major variants
Disgust	9	
	9+16+15, 26	
	9+17	
	10*	
	10*+16+25, 26	
	10+17	
Anger	4+5*+7+10*+22+23+25,26	
	4+5*+7+l0*+23+25,26	
	4+5*+7+23+25, 26	
	4+5*+7+17+23	
	4+5*+7+17+24	
	4+5*+7+23	
	4+5*+7+24	

establish a very intuitive and natural way of communicating emotions, and also to provide a positive influence on the user by means of evocative associations.

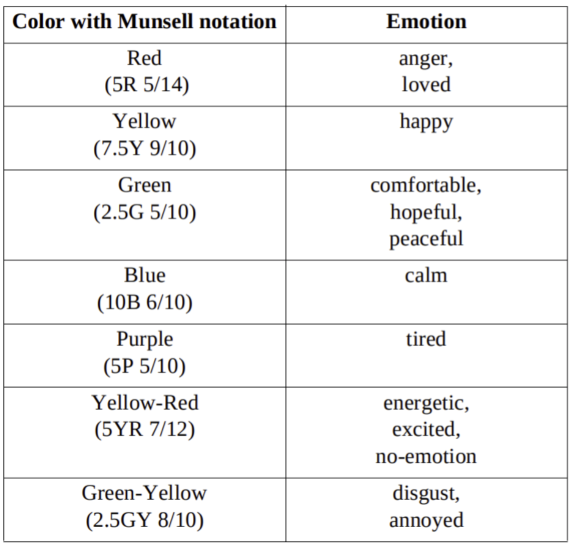
In order to use colors to stimulate a certain positive feeling, such as calm, energy, happiness we first need to have a semantic mapping between colors and emotions. Plenty of experiments have been done in order to find a precise and reliable mapping, and all have provided almost the same results in terms of color-meaning. To achieve a simple and affordable model, the proposed method relies on Naz Kaya research [21] whose results are summarized in table 5.

Once found a good emotion-color mapping, an emphatic strategy that selects a specific evocative color depending on the particular recognized emotion can be adopted.

For example when the robot perceive that the user is in trouble or is afraid, lights can automatically switched to green color or green shadows in order to recall a sense of peace and create a more comfortable environment. Indeed color green is usually associated to the most positive emotions such as hopeful, peaceful and satisfaction. The same applies for blue color that is typically associated to calm and relax emotional states, and as for green color, can be useful to contrast negative emotions such as Angry. Studies have proven that yellow color is able to evoke happy and joy emotional states, therefore yellow lights can be set whenever the robot recognize that the user can be sad or even happy and surprise, in order to emphasize and agree with these positive emotions. Strong colors such as red are usually associated to anger, aggressive and very intense emotions and mixed with yellow shadows can evoke active, energetic and powerful motivational states. Then it is reasonable to activate Yellow-Red shadows when user is recognize to be sad and demotivated.

7. Conclusion

In this work, we address the problem of Facial Emotion Recognition by introducing an end-to-end trainable deep



Color with Munsell notation	Emotion
Red (5R 5/14)	anger, loved
Yellow (7.5Y 9/10)	happy
Green (2.5G 5/10)	comfortable, hopeful, peaceful
Blue (10B 6/10)	calm
Purple (5P 5/10)	tired
Yellow-Red (5YR 7/12)	energetic, excited, no-emotion
Green-Yellow (2.5GY 8/10)	disgust, annoyed

Figure 5: Naz Kaya emotions-colors mapping.

learning model which can reasoning on both spatial and temporal features of facial expressions. Results have shown as the proposed model can effectively learn to distinguish between most basic emotions. Furthermore, a method for setting RGB lights components according to recognized user emotions is presented. In particular an emphatic strategy that exploits the proposed emotion recognition system to identify user emotions and accordingly monitoring ambient colored lights can be used to improve the quality of human-robot interactions.

In the future, the proposed strategy can be developed and validated in a social robot like Pepper from SoftBank Robotics [22], by monitoring colors of its body LEDs. They are placed in the chest, eyes, shoulders and ears allowing for a more friendly and engaging interaction. Another possible direction of future work could explore how this emphatic model can become adaptive to the user preferences, in such a way the robot can learn the impact of the different colors in a particular user, depending on his previous reactions.

7.1. Ethical Impacts

The system presented in this paper has a wide field of employment. It's aim is to detect and point out a person's emotion in a very simple and informative way through colors. Applications of such systems can be useful when the environment needs to intelligently adapt to the user, i.e. changing the light color in response to people mood in a room full of music stimuli. But we also acknowledge the possibility of misuse. Indeed, emotions are a fundamental part of peoples live and thus are private. Advancing the state of the art in this field also means exposing every human inner self to the world. As for most of the research the pros and cons must be weighted and evaluated considering every possible use case scenario. We believe that our work does not hold enough practical ground to be misused by third actors, but we are still concerned with this possibility.

References

[1] S. Russo, S. Illari, R. Avanzato, C. Napoli, Reducing the psychological burden of isolated oncological patients by means of decision trees, volume 2768, 2020, pp. 46–53.

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

[3] N. A. Nijdam, Mapping emotion to color, Book Mapping emotion to color (2009) 2–9.

[4] H.-W. Ng, V. D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: Proceedings of the 2015 ACM on international conference on multimodal interaction, 2015, pp. 443–449.

[5] M. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, T. Shimamura, Facial emotion recognition using transfer learning in the deep cnn, Electronics 10 (2021) 1036.

[6] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Acted facial expressions in the wild database, Australian National University, Canberra, Australia, Technical Report TR-CS-11 2 (2011) 1.

[7] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2983–2991.

[8] J. Haddad, O. Lézoray, P. Hamel, 3d-cnn for facial emotion recognition in videos, in: International Symposium on Visual Computing, Springer, 2020, pp. 298–309.

[9] W.-S. Chu, F. De la Torre, J. F. Cohn, Learning spatial and temporal cues for multi-label facial action unit detection, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 25–32.

[10] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using cnn-rnn and c3d hybrid networks, in: Proceedings of the 18th ACM international conference on multimodal interaction, 2016, pp. 445–450.

[11] D. H. Kim, W. J. Baddar, J. Jang, Y. M. Ro, Multi-objective based spatio-temporal feature representation learning robust to expression intensity varia-

tions for facial expression recognition, IEEE Transactions on Affective Computing 10 (2017) 223–236.

[12] F. Jimenez, T. Ando, M. Kanoh, T. Nakamura, Psychological effects of a synchronously reliant agent on human beings, Journal of Advanced Computational Intelligence Vol 17 (2013).

[13] J. Feldmaier, T. Marmat, J. Kuhn, K. Diepold, Evaluation of a rgb-led-based emotion display for affective agents, arXiv preprint arXiv:1612.07303 (2016).

[14] R. Avanzato, F. Beritelli, M. Russo, S. Russo, M. Vaccaro, Yolov3-based mask and face recognition algorithm for individual protection applications, volume 2768, 2020, pp. 41–45.

[15] S. Russo, C. Napoli, A comprehensive solution for psychological treatment and therapeutic path planning based on knowledge base and expertise sharing, volume 2472, 2019, pp. 41–47.

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: 2010 ieee computer society conference on computer vision and pattern recognition-workshops, IEEE, 2010, pp. 94–101.

[17] I. Leite, G. Castellano, A. Pereira, C. Martinho, A. Paiva, Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings, in: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, 2012, pp. 367–374.

[18] C. Napoli, G. Pappalardo, E. Tramontana, Using modularity metrics to assist move method refactoring of large systems, 2013, pp. 529–534. doi:10.1109/CISIS.2013.96.

[19] M. Nalin, L. Bergamini, A. Giusti, I. Baroni, A. Sanna, Children's perception of a robotic companion in a mildly constrained setting, in: IEEE/ACM human-robot interaction 2011 conference (robots with children workshop) proceedings, Citeseer, 2011.

[20] M. Wozniak, D. Polap, G. Borowik, C. Napoli, A first attempt to cloud-based user verification in distributed system, in: 2015 Asia-Pacific Conference on Computer Aided System Engineering, IEEE, 2015, pp. 226–231.

[21] N. Kaya, H. H. Epps, Relationship between color and emotion: A study of college students, College student journal 38 (2004) 396–405.

[22] A. K. Pandey, R. Gelin, A mass-produced sociable humanoid robot: Pepper: The first machine of its kind, IEEE Robotics & Automation Magazine 25 (2018) 40–48.