Research Article

Empirical Studies of the Arts
1–27
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02762374231163481
journals.sagepub.com/home/art
SAGE

Detection of Emotions in Artworks Using a Convolutional Neural Network Trained on Non-Artistic Images: A Methodology to Reduce the Cross-Depiction Problem

César González-Martín[1], Miguel Carrasco[2], and Thomas Gustavo Wachter Wielandt[3]

Abstract

This research is framed within the study of automatic recognition of emotions in artworks, proposing a methodology to improve performance in detecting emotions when a network is trained with an image type different from the entry type, which is known as the cross-depiction problem. To achieve this, we used the QuickShift algorithm, which simplifies images' resources, and applied it to the Open Affective Standardized Image (OASIS) dataset as well as the WikiArt Emotion dataset. Both datasets are also unified under a binary emotional system. Subsequently, a model was trained based on a convolutional neural network using OASIS as a learning base, in order to then be applied on the WikiArt Emotion dataset. The results show an improvement in the general prediction performance when applying QuickShift (73% overall). However, we can observe that artistic style influences the results, with minimalist art being incompatible with the methodology proposed.

[1] Department of Specific Didactics, University of Cordoba, Cordoba, Spain
[2] Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Chile
[3] Department of Computer Science, Universidad Adolfo Ibáñez, Santiago, Chile

Corresponding Author:
César González-Martín, Department of Specific Didactics, University of Cordoba, San Alberto Magno, Cordoba 14071, Spain.
Email: cesar.gonzalez@uco.es

Keywords
emotion, art, cross-depiction problem, deep learning, QuickShift

Introduction

Images have the capacity to send a message with a given emotional state. Depending on their goal, certain denotative elements are used in image creation, called low-level or hand-crafted features (Wang, Han, & Jin, 2019), which will compose the connotative or high-level message (Machajdik & Hanbury, 2010) which contains the message and the emotion. The analysis can focus on the color (González-Martín, Carrasco, & Oviedo, 2022; Ou, Luo, Woodcock, & Wright, 2004a, 2004b, 2004c; Tadi Bani & Fekri-Ershad, 2019), texture (Liu, Jiang, Pei, & Liu, 2018; Liu & Pei, 2018; Shao, Zhou, Cheng, Diao, & Zhang, 2015), forms (Lu et al., 2012), and detection of the objects which comprise the image (Campos, Jou, & Giró-i-Nieto, 2017; Sowmyayani & Rani, 2022; Zhang, Liu, Chen, Ye, & Wang, 2022). Other studies have attempted to go deeper into extracting other aspects which aid the transmission of emotions such as balance, emphasis, harmony, variety, gradation, and movement (Zhao et al., 2014) or the aesthetics (Fekete et al., 2022; Joshi et al., 2011).

Studying this is complex, though, since each denotative element has a meaning, and this can change with the interrelation with other elements in the representation space. This process is also determined by the concrete sociocultural context of the subject (Lim, 2016; Russell, 2017). To approach the analysis of emotions in fixed and moving images, the literature has presented various different methodologies (Wang et al., 2019), as well as diverse study objects, including facial expression analysis (Li & Xu, 2020), body language (Sapinski, Kaminska, Pelikant, & Anbarjafari, 2019), head position (Samanta & Guha, 2021), the context for where the scene is located (Kosti, Alvarez, Recasens, & Lapedriza, 2020; Mao, Zhu, Rao, Jia, & Luo, 2019) or the descriptive components of the image (Yanulevskaya et al., 2008; Zhang, He, & Lu, 2020).

Within the analysis of static images, art has been a field for experimentation and analysis regarding emotions. Thus, some studies have used abstract art (Chiarella et al., 2022; Sartori et al., 2015; Yanulevskaya et al., 2012; Zhang et al., 2011; Zhao et al., 2014), art styles including Oriental art (Hung, 2018), cubist art (Ginosar, Haas, Brown, & Malik, 2014), figurative art (Hagtvedt, Patrick, & Hagtvedt, 2008; Huang, Huang, & Kuo, 2010), or artworks from various cultures (Stamatopoulou & Cupchik, 2017), among others. Similarly, in this field, experiments have been done combining different resources apart from image analysis, such as using the title of the work or the author (Chamberlain, Mullin, Scheerlinck, & Wagemans, 2018; Chiarella et al., 2022; Huang, Bridge, Kemp, & Parker, 2011; Sartori et al., 2015; Tashu, Hajiyeva, & Horvath, 2021) to better define emotion.

As we can see, analyzing images and emotions is rich and complex. The literature presents an extensive list of tools and approaches for understanding emotions in different types of images. However, each of them is specific to the problem and the type of object and/or scene. To understand this problem, suppose you look at a photograph of a dog and ask a child to draw a picture of a dog. In both cases, we know that it corresponds to the same symbol and object, in this case, a dog. However, is it possible to determine the differences between the two objects through a computational tool? This would imply that there is some kind of characteristic or property that allows us to extract certain underlying symbols in the images and drawings (Huang, Wang, & Bai, 2021). The same problem occurs in the case of emotions when trying to understand and connect emotions in images that come from different typologies, styles, or formats. Thus, our proposal intends to contribute a methodological proposal to this field where a tool based on a deep learning architecture allows for inferring emotions from a set of artworks, even when the network has been trained on a set of photographs with real scenes (not artistic ones); that is, on a dataset which is unknown by the network. Formally, this problem is called the cross-depiction problem, referring to the difficulty which automatic learning models have in generalizing about data which are not exactly similar to those on which it was previously trained (Hall, Cai, Wu, & Corradi, 2015; Zhu et al., 2022). This problem has seen more studies in recent years to facilitate learning about different domains (Dash, Chitlangia, Ahuja, & Srinivasan, 2022) and it supposes a limitation in the field of deep learning.

Problem Focus

The problem addressed in this research has started to be studied in greater depth in different areas of machine learning. This is because many of the technologies based on deep learning have achieved excellent results in specific domains (semantic segmentation, object identification, emotion analysis, etc.) but also in the field of machine learning (Choi & Yun, 2020; Poterek, Herrault, Skupinski, & Sheeren, 2020; Renò et al., 2020; Sowmyayani & Rani, 2022). However, if you want to use a deep learning model in other contexts (not the one it was trained on), its performance drops rapidly (Huang et al., 2021; Sun & Saenko, 2016; Zhu et al., 2022). To deal with this limitation, we seek to understand whether it is possible to improve performance by modifying both the training and test images so that there is a similarity between the two sets.

In particular, this study uses two different image datasets: First, "Open Affective Scandalized Image Set" (OASIS), created as an open option to contribute to studying the relationship between images and emotions (Kurdi, Lozano, & Banaji, 2017) composed of 900 photographs without stylistic categorization will be on display for the training; and second, WikiArt Emotion, which will be the unknown input to the network, has 4,105 paintings extracted from WikiArt.org, falling into four major Western art styles: Renaissance Art, Post-Renaissance Art, Modern Art, and

Contemporary Art, with 22 style categories, and 20 emotions ranging from gratitude to sorrow, passing through angst and fear (Mohammad & Kiritchenko, 2018) (Figure 1).

Both databases provide the basis for answering our research question and have been selected for three reasons. First, these databases have a robust emotional model that has been widely validated by the community. Second, although they have different emotional models, previous experiments reported in the literature have managed to define a compatibility mechanism where the emotions of a continuous model (DES) are compatible with those of a discrete model (CES) through a process of emotion binarization. This last part is key to designing a model that can be applied to another source of information because without it, it would not be feasible to assign one model to another. Finally, the chosen datasets have different representational styles constituted by the denotative elements, which obfuscates the automatic detection of emotions. To address this limitation, we seek to understand whether it is possible to improve performance by modifying both training and test images so that there is a similarity between the two sets, through a process that modifies the compositional elements of the image, harmonizing the visual criteria of the sample.

In the following section, we analyze in depth the different tools that have been implemented for emotion recognition from a computational perspective, considering the emotional models on which they are built. Then, we detail the methodology applied in the project and explain all the necessary steps for the experimentation

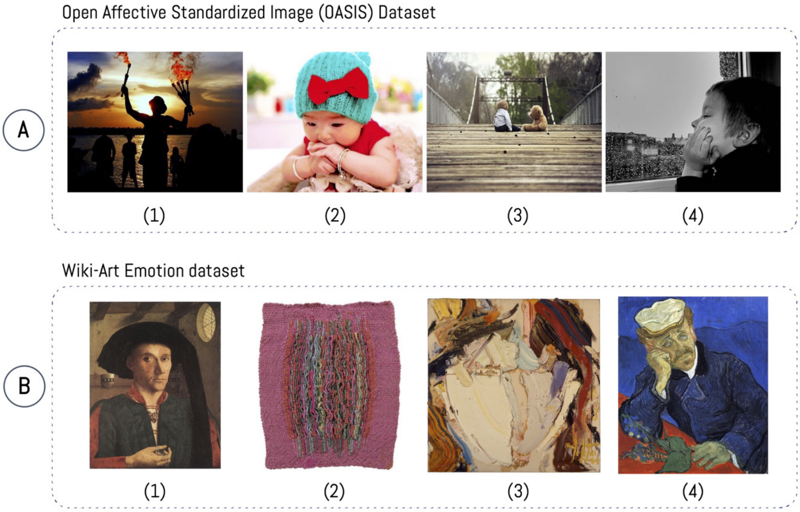

Figure 1. (A) Open Affective Standardized Image (OASIS) dataset: (1) celebration #2, (2) baby #9, (3) baby #10, and (4) bored pose #6 (the symbol # corresponds to a correlative image of that category). (B) WikiArt Emotions Dataset example: (1) Renassaince, (2) contemporary, (3) modern art, and (4) post-Renassaince.

and analysis phase. Finally, in the discussion and conclusion, we present the results obtained and the limitations of our experimental phase, which opens possible avenues for further research.

Recognizing Emotions: Affective Computation

The field of affective computation has roots going back more than three decades (Minsky, 1986; Picard, 1995). However, it is only in the last decade that it has been possible to advance in a deeper comprehension and formulation about emotional inference based on digital images (Chen, Zhang, & Allebach, 2015; Kim, Lee, & Provost, 2013; Ram et al., 2020; Yao et al., 2020; Zhao et al., 2014); what we today know as affective computing. This has been achieved thanks to major advances in computational techniques based on deep learning architectures, which have aided in advancing automatic comprehension of emotions. Zhao et al. (2022) define this task as the "Content Analysis of Affective Images," which is constructed on a dataset of previously cataloged images with an emotional model. Thus, via learning processes, it has been possible to build increasingly robust emotional models.

There are various proposals for models and categorization (Ortony, 2022). From a psychological standpoint, the main models used have been categorical emotion states (CES), which considers the emotions of fear, disgust, and sadness, and binarizes into positive–negative, following Ekman's six basic emotions, and Dimensional Emotion Space (DES) (Zhao et al., 2022), which uses valence variables, arousal, and dominance to describe the emotional intensity, among others. Arousal defines emotional intensity, valence defines the emotional type, and dominance is defined in relation to submission and control. The final variable is generally not used, since the images do not present relevant information about this variable. Thus, the most applied factor is normally the valence–arousal relation (Xu et al., 2013). Both emotional models respond to a subjective scale.

From a computational perspective, emotional analysis has been approached from different areas to try to understand how emotions are constituted. This can include using text analysis (Alswaidan & Menai, 2020; Batbaatar, Li, & Ryu, 2019; Gupta, Roy, Batra, & Dubey, 2021; Sundaram, Ahmed, Muqtadeer, & Reddy, 2021), speech (Kumar, Jain, Raman, Roy, & Iwamura, 2021; Li & Xu, 2020; Van, Nguyen, & Le, 2022; Zhou, Liang, Gu, Yin, & Yao, 2022; Zhang & Xue, 2021a, 2021b), music (Du, Li, & Gao, 2020; Putkinen et al., 2021; Xu, Xu, & Zhang, 2021), or a combination of multiple media (Shen, Zheng, & Wang, 2021), among others.

Centered on the image field, a fundamental requirement is to know about the representative labels of an emotional model (Bradley & Lang, 2007). One of the most often used is International Affective Picture System (IAPS), with 1,182 images tagged with DES (Lang, 1999). Under this model, we can find more current datasets with larger sample numbers, including ArtPhoto (Machajdik & Hanbury, 2010), Museum of Modern and Contemporary Art of Trento and Rovereto (MART)

(Alameda-Pineda, Ricci, Yan, & Sebe, 2016; Yanulevskaya et al., 2012), devArt (Alameda-Pineda et al., 2016), or ArtEmis (Achlioptas, Ovsjanikov, Haydarov, Elhoseiny, & Guibas, 2021). It is also possible to find datasets with a larger size, such as those described in Image-Emotion-Social-Net (IESN) with over a million images labeled under the CES and DES model (Zhao et al., 2016), and the Twitter for Sentiment Analysis Dataset (T4SA) with 1.5 million images categorized under a positive-negative-neutral model (Vadicamo et al., 2017). However, for our study, we have not used these databases, since in the case of IESN, the number of evaluators is very limited, and in T4SA the labeling process has no human intervention.

The major leap in emotional inference tools has been the rise of convolutional neural networks (CNNs), which have made it possible to face the problem of analysis and interpretation with better performance than previously mentioned techniques (Krizhevsky, Sutskever, & Hinton, 2017). The main advantage of CNNs is their capacity to extract underlying patterns from data without human intervention. The CNN itself extracts the characteristics allowing for adequate training in a supervised learning process, that is, where problem categories or classes are previously known. Thus, the CNN can separate non-linear and correlated characteristics about different data types. The applications covered by CNN are very broad and diverse. In the emotion detection area, some of them have focused on both the form and description of color (Kahou et al., 2015; Liu, Sun, Li, & Iida, 2020; Poterek et al., 2020; Razavian, Azizpour, Sullivan, & Carlsson, 2014; Renò et al., 2020; Wang & Deng, 2018; Westlake, Cai, & Hall, 2016).

The first studies with CNN used in emotion detection were those by Kim et al. (2013) and Razavian et al. (2014). In both cases, the authors used a network originally proposed by Krizhevsky et al. (2017) with some modifications in the network structure and the data type. In later years, research was centered on faces as a means of recognizing emotions. This area includes the studies by Kahou et al. (2015), Kollias, Marandianos, Raouzaiou, and Stafylopatis (2015), and Wei et al. (2017). Despite these advances, recent studies have focused not only on the face, but also on other aspects contained in the images, particularly color, forms, and composition (Elliot, 2015; He, Qi, & Zaretzki, 2015; Takada, Wang, & Yamasaki, 2021).

Emotion arises from observers' own experiences, which makes it essential to interpret the objects, forms, and colors present in images. This is true regardless of the medium, whether they are abstract images, artistic images, or digital photographs. These facts have lead to some researchers integrating both object information and image background (Kim, Kim, Kim, & Lee, 2017; Priya & Udayan, 2020). Although the color is a low-level characteristic, it has a relevant relationship with both the object and the meaning expressed in the image (González-Martín et al., 2022). Other researchers have proposed a combination of multiple CNNs to detect objects at a high semantic level, also considering texture and aesthetics (Lu, Lin, Jin, Yang, & Wang, 2014; Rao, Xu, & Xu, 2018).

As we can see, research in this field is quite broad and has sought to incorporate new relations and elements to improve performance for inference models in a similar way to the process which occurs in human beings. However, one of the main restrictions is that models are normally evaluated on the same types or data for which they were trained. That is, a model trained on photographic images is not evaluated on abstract works. This problem has been formally analyzed in recent years, allowing for definitions of both the dominion of the problem and the internal construction of deep learning algorithms (Dash et al., 2022). In our case, the transformation is proposed in the stage of data entry into the network. What we seek is to reduce the discrepancy between models from different domains (Huang et al., 2021; Zhu et al., 2022). Unlike other studies, our research proposes to understand the domain change of emotions generated from images of different domains, and not to relate compression of the objects contained in the images.

Methodology

To carry out this study, we propose the following stages: (1) simplification of the image dataset via the QuickShift algorithm (Vedaldi & Soatto, 2008); (2) emotion extraction with deep learning techniques; and (3) learning evaluation on a dataset of artworks associated with a given emotion. We will now detail each phase of the described process in Figure 2.

Step (I) Data Preparation

The datasets used in this research are OASIS and WikiArt Emotion. Both datasets are transformed by simplifying the images with the QuickShift algorithm, which allows for grouping pixels with similar properties via a clustering process. The algorithm organizes all the data points in a tree, where the parents in the tree are the closest neighbors in the characteristics space, which are organized around increasing density (Vedaldi & Soatto, 2008). As

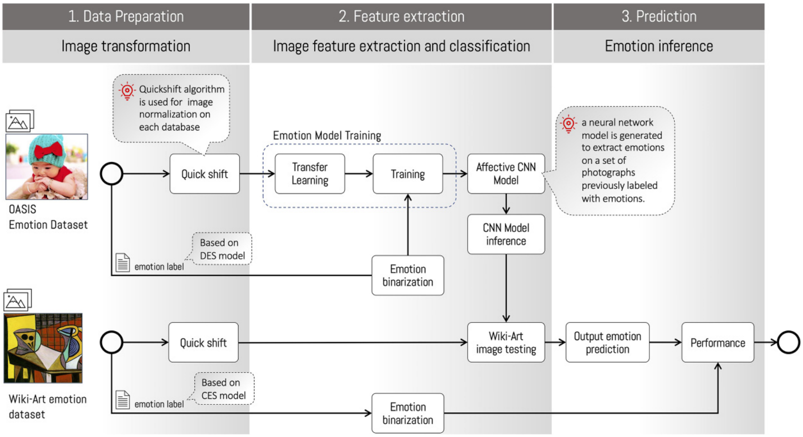

Figure 2. Diagram of the learning process and emotion prediction in artworks.

we commented, this pre-processing is applied both to training images (digital photographs of OASIS) and the artworks (WikiArt Emotion). As a result, by applying the algorithm to both datasets, the images share similar visual resources (example in Figure 3). To evaluate the behavior in data preparation, we did different experiments with and without the use of the QuickShift algorithm in order to know whether using a clustering algorithm generated a better final performance. Subsequently, we analyzed the behavior of the clustering parameter in the QuickShift algorithm, in order to study its sensitivity with the classification process (see experiment 1). This process is performed with the same parameter for both datasets. This provides the basis for the next stage of the classification and feature extraction process.

Step (II) *Feature Extraction*

This stage performs two independent processes, which make it possible to create a binary contrast between the emotions contained in the images of both datasets, since

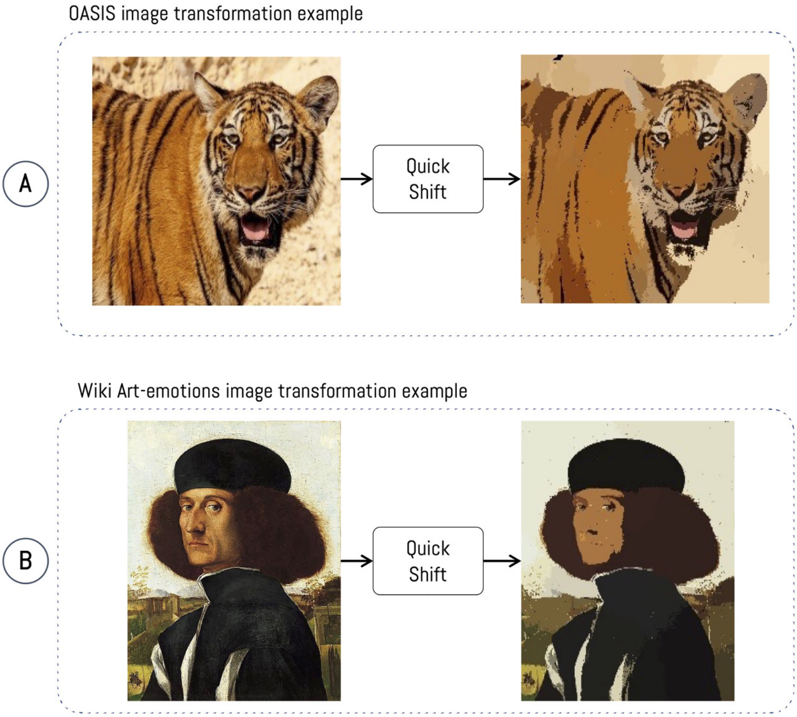
OASIS image transformation example
A
Quick Shift
Wiki Art-emotions image transformation example
B
Quick Shift

Figure 3. Application of the QuickShift algorithm on the Open Affective Standardized Image (OASIS) and WikiArt databases (a) OASIS Tiger #2 (tiger2_2.jpg), (b) portrait of a Venetian nobleman, Vittore Carpaccio (1455–1526, Norton Simon Museum).

they have different emotional categories (apart from the content type and style of the images), as well as making it possible to perform the training process.

Creating binary contrasts and dichotomies for feelings is a simplification of the complexity and depth of emotions which humans can express. However, it has been an accepted approximation in other studies, mainly due to the high number of emotions described by other emotional models (Achlioptas et al., 2021; Campos et al., 2017).

Step (II.1) Emotional Binarization of the OASIS Dataset. OASIS dataset is composed of 900 photographs, which are categorized into measurements of arousal, valence, and dominance (DES scale). Our methodology uses the proposal of Lu et al. (2016) which is based solely on the valence descriptor since it allows for defining emotion between two states: positive and negative. Arousal is thus unnecessary since it only defines the emotional intensity. Furthermore, dominance will not be used since the images do not present relevant information about this variable. Therefore, in line with this author (Lu et al., 2016), if valence is above the median (4.33) the image is categorized as positive, and in the contrary case, as negative (Figure 4).

After this transformation, the sample obtained from OASIS for our study was binarised with 403 images as negative and 497 as positive, obtaining a balanced result for the training process. This procedure is relevant to avoid learning biases in a given class, improving the classifier performance (Krawczyk, 2016).

Step (II.2) Emotional Binarization of the WikiArt Emotion Dataset. The WikiArt Emotion dataset contains 20 emotional discrete states (CES). It, therefore, requires binarization into positive and negative states, allowing us to prove that the model training is correct. To this end, we have used the conversion scheme proposed by Mohammad and

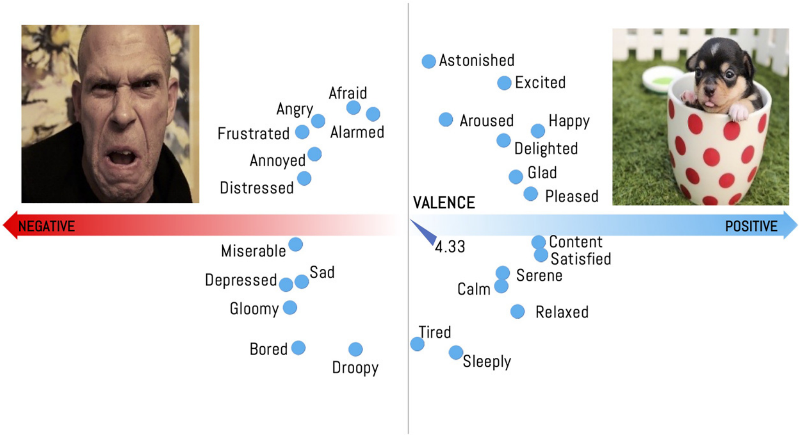

Figure 4. Relationship between Categorical Emotion States (CES) and Dimensional Emotion Space (DES) based on "The Circumplex Model of Affect" diagram by Russell (1980).

Kiritchenko (2018) (see Table 1). Mixed emotions and others have not been considered in conversion because they are not binarised, so these images have not been included in our analysis.

Step (II.3) Model Training. The model training stage consists of designing a CNN which allows for inferring the emotions of a previously labeled dataset, in this case, on the OASIS dataset. This allows the CNN to extract underlying patterns from the data which describe their emotions. Sadly, the image set is very limited for performing the classification task. For this reason, we have used a pre-trained neural network, specifically the Xception network (Chollet, 2017). This network was trained for classifying the Imagenet skill, having been previously trained on 1 million images categorized into 1000 different classes (Deng et al., 2009; Russakovsky et al., 2015). To adapt this network to the problem of extracting feelings, the upper layers and final layers of the network were re-trained. This was done in order for the entrance and exit weights to be adapted to the problem, although the previously trained deep layers' weights were maintained. We also changed the number of neurons in the exit layer from 1000 to 1, and its activation function was changed to the "sigmoid" function in order to achieve a single numerical value in the range of 0 to 1, which is consistent with the binary exit logic. Subsequently, the learning acquired with OASIS was applied to

Table 1. Conversion Between Positive, Negative, and Other and Mixed Emotions Based on the Study by Mohammad and Kiritchenko (2018).

Polarity	Emotion
Positive	Gratitude, thankfulness
	Happiness, calmness, pleasure
	Humility, modesty, unpretentiousness, simplicity
	Love, affection
	Optimist, hopefulness, confidence
	Trust, admiration, respect, dignity, honor
Negative	Anger, annoyance, rage
	Arrogance, vanity, hubris, conceit
	Disgust, dislike, indifference, hate
	Fear, anxiety
	Pessimism, cynicism
	Regret, guilt, remorse
	Sadness, loneliness, grief
	Shame, humiliation, disgrace
Other or Mixed	Agreeableness, acceptance, submission, or compliance
	Anticipation, interest, curiosity, suspicion, or vigilance
	Disagreeableness, defiance, conflict, or strife
	Surprise, surrealism, amazement, or confusion
	Shyness, self-consciousness, reserve, reticence, shy, neutral

the WikiArt Emotions Dataset via an inference of the model. Thus, while we use 100% of the OASIS images, we use a network structure based on a larger set of images (Russakovsky et al., 2015), achieving a greater capacity for learning the input and greater flexibility when using another test dataset.

Step (III) Prediction

Prediction consists of classifying an emotion based on a given image from the WikiArt Emotions data set (artworks), via the training generated with the OASIS dataset (photographs). We should remember that the group of images used in the artworks from WikiArt Emotion (Table 1) is larger than the emotions used in OASIS. To determine whether the prediction is correct, we compare the result obtained from the neural network with respect to the real class of that image. To carry out this process, we began by binarizing the emotions from WikiArt images (step II.2) followed by evaluating whether the CNN prediction aligns with the binarization of the entry image (WikiArt Emotion). Thus, the proposal allows us to extract emotions associated with a probability vector for each of the WikiArt Emotion images. In this way, if the network exit value surpasses the 0.5 threshold, the image is classified as positive; otherwise, it is classified as negative. To evaluate the performance of the model, given that each image used in the evaluation phase has a pre-binarised emotional model, we only have to compare the prediction with the real value of the said image (performance stage).

Results and Experimental Tests

This section presents the results obtained with the proposed methodology regarding emotional prediction in artworks via training using non-artistic images (digital photographs from the OASIS dataset (Kurdi et al., 2017)). Two experiments have been done. The first is a performance evaluation of the QuickShift algorithm, while the second focuses on determining the optimal parameter to simplify the image in order to maximize emotional classification.

Experiment 1. QuickShift Algorithm Evaluation

To evaluate the impact of the QuickShift algorithm, we present three configurations from the two datasets used in our study (Table 2). As an evaluation measure, we have used the combined $F1$-score, because it combines the Precision and Recall measures into a single number. Precision (P) is a measure of the percentage of positive predictions out of the total number of items with a class defined as positive. On the other hand, Recall (R) corresponds to the percentage of correctly classified positive images out of the set of positive elements. These two indicators are combined through the $F1$-score indicator as $F1 = \frac{2 \cdot P \cdot R}{P+R}$. Recall that the problem has been defined as a binary classification (positive and negative emotions). For each class, the algorithm

Table 2. Performance of the Proposed Model With Various Evaluation Configurations.

Train	Test	Avg F1-score model OASIS dataset	Avg F1-score test WikiArt dataset
OASIS sans QuickShift	WikiArt sans QuickShift	94.5% ± 2%	62% ± 5%
OASIS + QuickShift	WikiArt sans QuickShift	89% ± 5%	64% ± 6%
OASIS + QuickShift	WikiArt + QuickShift	89% ± 5%	73% ± 2%

must be able to differentiate between both emotions, the best performance is achieved when F1-score is 100%.

Results indicate that by applying training with images and then evaluating this learning on unfiltered artworks we obtained an F1-score test of 62% for the whole image set. These results show low performance in the evaluation of images which were unfamiliar to the network during the performance process, which is known as the cross-depiction problem. To deal with this problem, Huang et al. (2021) propose to reduce the discrepancy between image domains through the fusion of convolutional networks via style transfer. The experiments conducted in that research allow the comparison between four domains: real, paint, cartoon, and sketch. For our problem, we have solved the discrepancy more directly using the QuickShift algorithm, which allows us to simplify the visual resources of the images and thus unify styles across the sample categories.

When the QuickShif algorithm is applied to the OASIS images, we obtained a lower value in the training process since the network could not correctly interpret the emotion with the training image. However, we noted a slight improvement in the evaluation phase, obtaining an F1-score performance of 64%. Finally, when applying QuickShift to both datasets (OASIS and WikiArt), we obtained a significant improvement by obtaining on average an F1-score of 73% ± 2%, in comparison with 62% ± 5% without QuickShift on the artwork images (of a different type on average). These results show considerable mitigation on the cross-depiction problem, as long as the test images and the tests are conditioned via QuickShift with the same parameter. These results are consistent with those described by Huang et al. (2021) in obtaining a performance of over 70% for solving the cross-depiction problem.

Experiment 2. QuickShift Parameter Sensitivity Analysis

In the following analyses, we will perform exhaustive tests on the proposed method, using the basis that both datasets have been transformed via the QuickShift algorithm, and analyzing the behavior of the parameter-defining image simplification (both training and test).

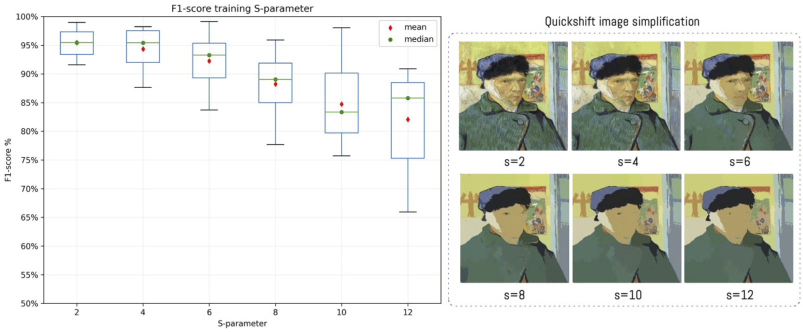

Figure 5. Box-plot graph of the training model performance according to the *s* parameter of the QuickShift algorithm which helps simplify images. The resulting image after successively increasing the *s*-parameter of the QuickShift algorithm.

As mentioned before, the training phase allows the network to make predictions about data which are unknown to the network. However, performance in this phase depends on the type of image used for training. As we see in Figure 5, the $F1$-score performance during the training phase varies according to the "s" parameter. This means that as the parameter increases, the number of colors in the image drops, resulting in a simplification of the components and structures in the resulting image. In the experimental phase, we have modified the parameter from 2 to 12 with an increase of two units. Values greater than 12 over-simplify the image, and have thus not been used. For each value of the s parameter, we have done 10 evaluations in order to record algorithm performance. We have thus observed that starting at $s > = 6$, algorithm dispersion begins to increase in the prediction phase, implying that error increases (Figure 6).

To visualize algorithmic behavior while taking all the parameter changes into account, we can observe that contemporary art has inferior performance compared with other art types (on average under 25%) (Figure 7). The other typologies have a performance varying between 68% and 77% on their $F1$-score, with the Renaissance type being slightly higher between them, but without any relevant statistical difference.

The artworks in the WikiArt Emotions dataset are distributed across four artistic styles. To analyze the behavior of the proposed methodology, we have studied algorithmic behavior according to this classification (see Figure 6). In each case, we observe that performance depended on the s-parameter. However, unlike the training process, as the parameter increases, performance rises, which stabilizes at $s = 10$. The performance also varies by artistic style. The lowest performance and the highest variability appear in contemporary art (mean $F1$-score $= 0.40$, $s = 6$), with the best performance being obtained in post-Renaissance art (mean $F1$-score $= 0.776$, $s = 12$). As in the model training process, each parameter was analyzed 10 times for each style. To

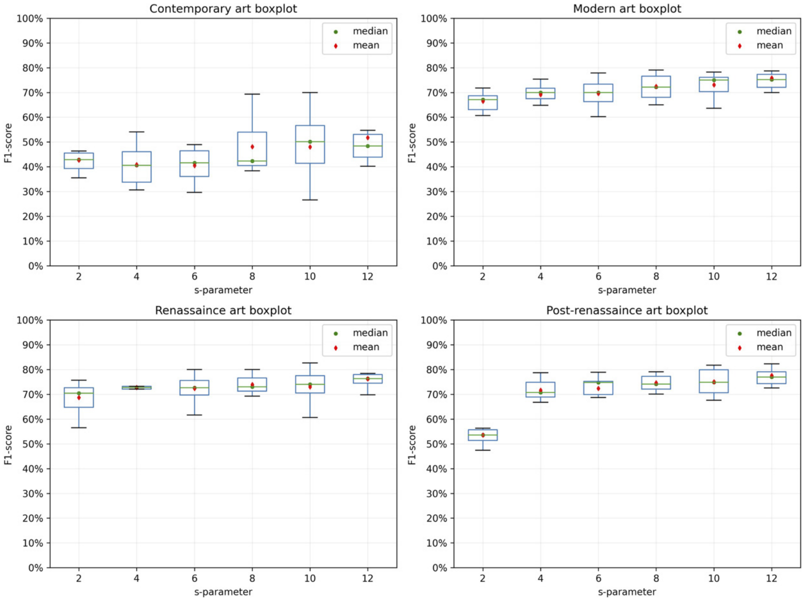

Figure 6. Boxplot graph of performance by the art style and QuickShift algorithm application.

visualize the behavior, the boxplot graph presents performance variations for each evaluation by type and parameter (Figure 6). Table 3 presents the same values as shown in Figure 6, but in addition, the standard deviation and averages are presented as the adjustment s-parameter increases.

Apart from contemporary art, the art styles have similar behavior, increasing their performance as image simplification rises. This is related to the artistic structure of the works contained in the contemporary art category, which are simpler compared to other art classifications. The effect generates image hyper-simplification so that the algorithm loses the capacity to find specific patterns in the elements composing the images (Figure 8). On the other hand, we have carried out a t-Welch test comparison between the four artistic styles of the WikiArt Emotion dataset. The results indicate no significant statistical difference between modern, post-Renassaince, and Renassaince art styles with $s = 12$. However, there is a difference between that styles and contemporary art (see Table 4).

Another way to visualize the parameter effect is to analyze the confidence interval of each parameter change by art type (see Figure 9). In this case, by combining all types, we can see that starting at s>=4 the typologies of modern art, Renaissance, and post-Renaissance present similar performance. However, in the case of contemporary art, performance is remarkably inferior, with a confidence interval increasing as the parameter increases. We can also clearly observe how the prediction of the model decreases in the training stage.

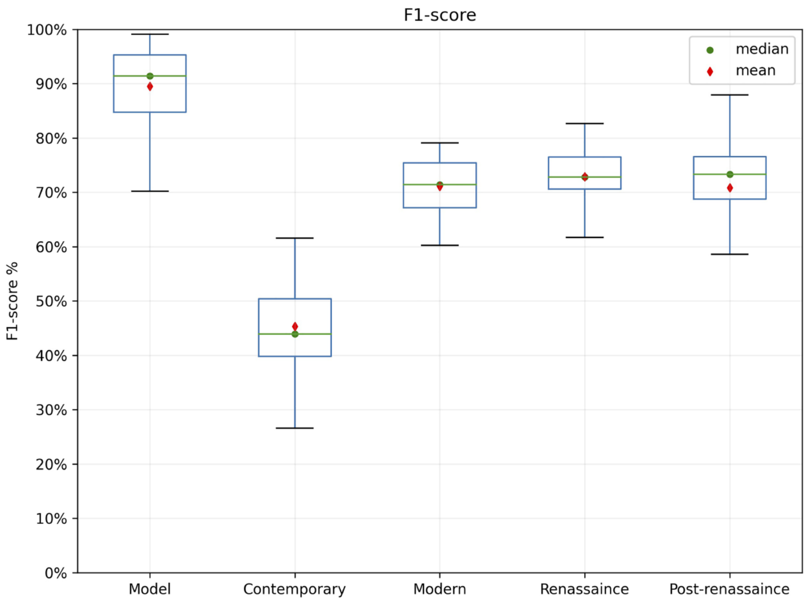

Figure 7. Global performance comparison by the art style and QuickShift algorithm application, considering the *s* parameter combination in the range [2–12].

Table 3. Average Performance and Standard Deviation According to Change in QuickShift s-Parameter, and According to Art Style.

	OASIS dataset		WikiArt dataset (used only for prediction)							
QS parameter	Model		Contemporary		Modern		Post-Renassaince		Renassaince	
s	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
2	95.43%	0.03	42.7%	0.08	66.4%	0.04	53.6%	0.04	68.7%	0.06
4	94.32%	0.04	40.9%	0.08	69.1%	0.04	71.7%	0.04	72.8%	0.04
6	92.26%	0.05	40.4%	0.07	69.6%	0.05	72.4%	0.06	72.4%	0.06
8	88.21%	0.06	48.1%	0.11	72.4%	0.05	74.7%	0.03	73.9%	0.04
10	84.74%	0.08	48.0%	0.13	73.1%	0.05	75.1%	0.05	73.0%	0.07
12	82.07%	0.09	51.8%	0.15	**75.8%**	0.06	**77.6%**	0.05	**76.3%**	0.04

In this way, the results suggest that image simplification, both from the training set (OASIS) and the test set (WikiArt) allow the model to increase its general prediction level. This is mainly because of reduced visual resources. However, if the elements comprising the image are oversimplified (contemporary art), performance will be affected.

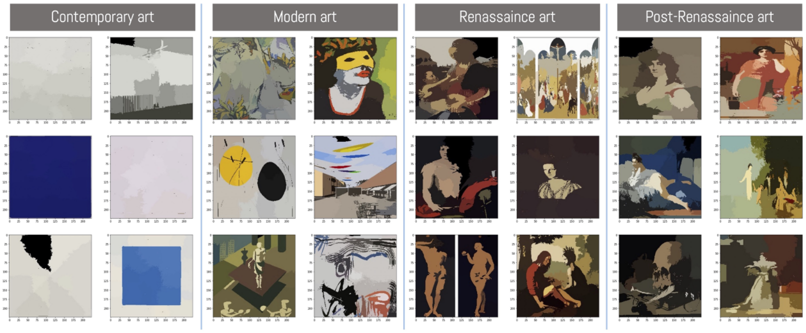

Figure 8. Incorrectly classified images by an art class. We can see that the elements present in contemporary art are oversimplified compared to other styles.

Table 4. t-Welch Test Between the Artistic Styles When the Parameter $s = 12$.

$s = 12$	Contemporary	Modern	Post-Renassaince	Renassaince
Contemporary	—	4.833*	5.274*	5.035*
Modern	−4.833*	—	0.773	0.188
Post-Renassaince	−5.274*	−0.773	—	−0.669
Renassaince	−5.035*	−0.188	0.669	—

*Significant difference.

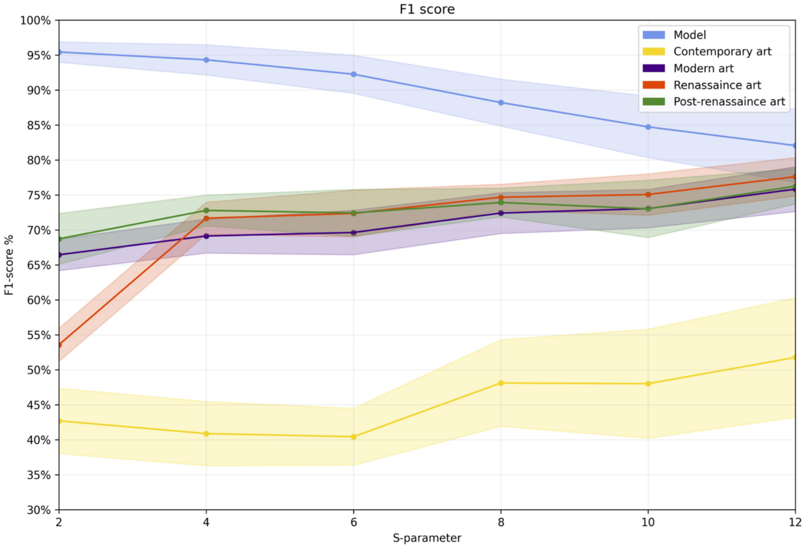

Figure 9. Evolution of the *s*-parameter in algorithmic performance. The shaded area corresponds to a 95% confidence interval.

Table 5. Comparative Performance of Our Method Versus Similar Techniques.

Study	Technique used	Cat. #*	Mtc.	Training data	Eval. Data	Results
Yanulevskaya	Manual	8	Acc	IAPS (photos)	IAPS	50% overall
Chen et al. (2015)	CNN	8	TP rate	Artphoto + own dataset	Photos	±70% overall
Campos et al. (2017)	CNN	2	Acc.	Twitter DeepSent dataset (photos)	Photos	83%
Lu et al. (2016)	Domain adaptation	2	Acc.	IAPS (photos)	Art	61%
Ours	CNN	2	Acc.	OASIS (photos)	Art (WikiArt)	77%* (73% overall)

Abbreviations: CNN = convolutional neural network; IAPS = International Affective Picture System; OASIS = Open Affective Standardized Image; TP = true positive; Acc = accuracy.
*Best performance.

Although it is not simple to compare with other algorithms described in the literature given the use of various databases, similar strategies have been reported with performances slightly inferior to ours (see Table 5). The novelty of this strategy thus lies in incorporating image simplification analysis via the QuickShift algorithm, which has helped increase performance in the prediction phase. However, this depends on the art style being used (Figure 9). The proposed methodology has also revealed that network performance can be raised, even while using the same neural network structure.

Discussion

For achieve the objectives proposed in this research, two experiments were required. In the first one, the performance of the QuickShift algorithm has been evaluated. The results obtained show that the best $F1$-score is produced by applying QuickShift with the same parameters on the two image datasets selected for this study with 73%, showing considerable mitigation of the cross-depiction problem.

Likewise, we needed to understand the optimal degree of simplification (parameter-s) that the algorithm can achieve per artistic style in the Art Emotion Wiki dataset, without losing the visual reference needed to recognize the emotion. Thus, in the second experiment carried out, the best performance is produced with Renaissance art (mean $F1$-score = 0.77, $s = 12$). On the other hand, the worst performance was with contemporary art, with an $F1$-score = 0.40 and $s = 6$. This is because the contemporary art category of the Art Emotion Wiki dataset is composed solely of minimalist artworks, which are abstract images, and therefore have a high level of simplification from their origin. The application of the QuickShift algorithm on this sample produces a visual hyper-simplification that prevents the recognition of emotions. With parameter s>=4, the modern art, Renaissance, and post-Renaissance typologies show similar performances. We have also found that if the s-parameter is greater than 12, the image becomes too schematic to recognize emotions with the proposed methodology.

As mentioned, modern art, Renaissance art, and post-renaissance art have similar behavior, while contemporary art shows a statistically significant difference. This has also been verified with the t-Welch test. The reason is the visual simplification when QuickShift is applied and the representation style of the images. Modern, Renaissance, and post-renaissance art have a similar level of figuration, while contemporary art, with minimalist artworks, has a higher level of abstraction in its image set, which produces these statistics results.

Conclusion

As we have seen, in recent years, affective computing has considerably increased its applications and study fields. However, it continues to present some problems and uncertainties in the processes (Landowska, 2019). On top of this, there is the problem presented by studying emotions and the number of variables participating in their construction, as well as analyzing and studying communication strategies via artworks to transmit emotions.

From a computational viewpoint, most of the tools which try to infer emotions do it on a single study object. For example, if we center on the artistic field when the object is painting, learning will not be applied to another art field. Our study has looked deeper into this problem, proposing a novel methodology which allows for learning emotions from a dataset with non-artistic samples and applying them in another domain, in this case art.

This has been done using a deep learning architecture on a database of photographic images without artistic classification, called OASIS (Kurdi et al., 2017), to then infer emotions about the artwork database called WikiArt Emotion. The method thus makes predictions about a set which is unknown to the CNN, as it was trained on another type of image and emotions in the training process, which is a novel phenomenon.

Via the presented methodology, we were able to achieve $F1$-accuracy above 70% in the task of classifying artworks in the sample used, except for contemporary art. The poor performance of contemporary art can be explained by the works contained in this category, which in the WikiArt Emotion Dataset include minimalist paintings. This type of artwork is characterized by elimination of aesthetics, subjectivity, and emotion by the artist (Best, 2006; Wolff, 2005), with only the “purest” elements remaining, and a predominance of orderly geometric shapes. These artworks are thus already simplified, so that after applying the QuickShift algorithm from our methodology they become oversimplified, providing very different results by comparison with other art categories.

The QuickShift algorithm also managed to reduce the cross-depiction problem produced when training a model on photographs and evaluating it with images from various art styles. Along the same lines, as shown in the results table (Table 2), the task of learning patterns in the images with QuickShift increases performance. This is probably due to the decreased visual elements in the image, a process which occurs in the first network layers, which facilitates learning and later emotional classification.

The proposed methodology is a novelty and can be significant in art-emotion relationship studies, from different areas of knowledge within the empirical aesthetics, allowing to extract emotions from an image without a previous emotional model, which opens wide research fields. Furthermore, we have analyzed how the image simplification through the QuickShift algorithm can be an efficient way to simplify the visual criteria of both the training set and the test set. As a result, it is possible to use this methodology for the automatic detection of emotional recognition independently of the artistic style applied in the creation of the artwork.

Limitations and Future Lines of Research

The results obtained present challenges and future action courses in the study field of emotions associated with artworks. As we have seen, when handling highly schematic work such as minimalist paintings, the proposed methodology is highly unsatisfactory given the over-simplification of the images after processing with the QuickShift algorithm. This presents limitations for automatic emotion recognition in images and artworks with few stimuli, opening possible future research lines on the limits of automatic emotion recognition.

Minimalist art is also characterized by its integration and interaction with the exposition space (Best, 2006), which is an impediment for the model presented in this study since two-dimensional images were used for training. Detecting emotions in three-dimensional space is an unknown challenge in computational analysis (Zhao et al., 2022).

Similarly, although performance improves with the QuickShift algorithm in the rest of the analyzed artistic styles, and helps mitigate the cross-depiction problem, simplification of the images, leading to their uniformity, causes a loss of aspects which enrich the communicative experience of the artworks, such as the rhythms in the brushstrokes, the shades of colors, subtle compositional elements, etc., which calls into question the information acquired with automated processes.

In this study, the use of emotions has been limited to a binary process of positive-negative. In future studies, we can contemplate enriching emotional nuances in artwork analysis.

Another limitation found in this study is the number of extant datasets which contain a significant number of images for automation processes with emotions associated with humans.

Data and Code Link

https://github.com/thomaswachterw/emotion-recognition-UAI.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by RRREMAKER. RRREMAKER MSC-RISE-H2020 project has received funding from the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie, proposal number: 101008060.

ORCID iD

César González-Martín https://orcid.org/0000-0002-9017-3196

References

Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., & Guibas, L. (2021). Artemis: Affective language for visual art. *Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11564–11574). Nashville, TN, USA, 2021. https://doi.org/10.1109/CVPR46437.2021.01140

Alameda-Pineda, X., Ricci, E., Yan, Y., & Sebe, N. (2016). Recognizing emotions from abstract paintings using non-linear matrix completion. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5240–5248). Las Vegas, NV, USA, 2016. https://doi.org/10.1109/CVPR.2016.566

Alswaidan, N., & Menai, M. E. B. (2020). A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, *62*(8), 2937–2987. https://doi.org/10.1007/s10115-020-01449-0

Batbaatar, E., Li, M., & Ryu, K. H. (2019). Semantic-Emotion neural network for emotion recognition from text. *Ieee Access*, *7*, 111866–111878. https://doi.org/10.1109/ACCESS.2019.2934529

Best, S. (2006). Minimalism, subjectivity, and aesthetics: Rethinking the anti-aesthetic tradition in late-modern art. *Journal of Visual Art Practice*, *5*(3), 127–142. https://doi.org/10.1386/jvap.5.3.127_1

Bradley, M. M., & Lang, P. J. (2007). Emotion and motivation. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (3rd ed, pp. 581–607). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511546396.025

Campos, V., Jou, B., & Giró-i-Nieto, X. (2017). From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image and Vision Computing*, *65*, 15–22. https://doi.org/10.1016/j.imavis.2017.01.011

Chamberlain, R., Mullin, C., Scheerlinck, B., & Wagemans, J. (2018). Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychology of Aesthetics, Creativity, and the Arts*, *12*(2), 177–192. https://doi.org/10.1037/aca0000136

Chen, M., Zhang, L., & Allebach, J. P. (2015). Learning deep features for image emotion classification. *IEEE International Conference on Image Processing (ICIP)* (pp. 4491–4495). Quebec City, QC, Canada, 2015. https://doi.org/10.1109/ICIP.2015.7351656

Chiarella, S. G., Torromino, G., Gagliardi, D. M., Rossi, D., Babiloni, F., & Cartocci, G. (2022). Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. *Computers in Human Behavior*, *137*(C). https://doi.org/10.1016/j.chb.2022.107406

Choi, H.-H., & Yun, B.-J. (2020). Deep learning-based computational color constancy with convoluted mixture of deep experts (CMoDE) fusion technique. *IEEE Access*, *8*, 188309–188320. https://doi.org/10.1109/ACCESS.2020.3030912

Chollet, F. (2017). *Xception: Deep Learning with Depthwise Separable Convolutions* (arXiv:1610.02357). arXiv. http://arxiv.org/abs/1610.02357

Dash, T., Chitlangia, S., Ahuja, A., & Srinivasan, A. (2022). A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, *12*(1), 1040. https://doi.org/10.1038/s41598-021-04590-0

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database, *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). Miami, FL, USA, 2009. https://doi.org/10.1109/CVPR.2009.5206848.

Du, P., Li, X., & Gao, Y. (2020). Dynamic Music emotion recognition based on CNN-BiLSTM. *IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)* (pp. 1372–1376). Chongqing, China, 2020. https://doi.org/10.1109/ITOEC49072.2020.9141729

Elliot, A. J. (2015). Color and psychological functioning: A review of theoretical and empirical work. *Frontiers in Psychology*, *6*, https://doi.org/10.3389/fpsyg.2015.00368

Fekete, A., Pelowski, M., Specker, E., Brieber, D., Rosenberg, R., & Leder, H. (2022). The Vienna Art Picture System (VAPS): A data set of 999 paintings and subjective ratings for art and aesthetics research. *Psychology of Aesthetics, Creativity, and the Arts*, Advance online publication. https://doi.org/10.1037/aca0000460

Ginosar, S., Haas, D., Brown, T., & Malik, J. (2014). Detecting People in Cubist Art. *ArXiv:1409.6235 [Cs]*. http://arxiv.org/abs/1409.6235

González-Martín, C., Carrasco, M., & Oviedo, G. (2022). *Analysis of the use of color and its emotional relationship in visual creations based on experiences during the context of the COVID-19 pandemic* (arXiv:2203.13770). arXiv. http://arxiv.org/abs/2203.13770

Gupta, P., Roy, I., Batra, G., & Dubey, A. K. (2021). Decoding Emotions in Text Using GloVe Embeddings. *International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 36–40). Greater Noida, India, 2021. https://doi.org/10.1109/ICCCIS51004.2021.9397132

Hagtvedt, H., Patrick, V. M., & Hagtvedt, R. (2008). The perception and evaluation of visual art. *Empirical Studies of the Arts*, *26*(2), 197–218. https://doi.org/10.2190/EM.26.2.d

Hall, P., Cai, H., Wu, Q., & Corradi, T. (2015). Cross-depiction problem: Recognition and synthesis of photographs and artwork. *Computational Visual Media*, *1*(2), 91–103. https://doi.org/10.1007/s41095-015-0017-1

He, L., Qi, H., & Zaretzki, R. (2015). Image color transfer to evoke different emotions based on color combinations. *Signal, Image and Video Processing*, *9*(8), 1965–1973. https://doi.org/10.1007/s11760-014-0691-y

Huang, K.-C., Huang, S.-Y., & Kuo, Y.-H. (2010). Emotion recognition based on a novel triangular facial feature extraction method. *The 2010 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–6). Barcelona, Spain, 2010. https://doi.org/10.1109/IJCNN.2010.5596374.

Huang, L., Wang, Y., & Bai, T. (2021). Recognizing art work image from natural type: A deep adaptive depiction fusion method. *The Visual Computer*, *37*(5), 1221–1232. https://doi.org/10.1007/s00371-020-01995-2

Huang, M., Bridge, H., Kemp, M., & Parker, A. (2011). Human cortical activity evoked by the assignment of authenticity when viewing works of art. *Frontiers in Human Neuroscience*, *5*, https://www.frontiersin.org/articles/10.3389/fnhum.2011.00134

Hung, C.-C. (2018). A study on a content-based image retrieval technique for Chinese paintings. *The Electronic Library*, *36*(1), 172–188. https://doi.org/10.1108/EL-10-2016-0219

Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q.-T., Wang, J., Li, J., & Luo, J. (2011). Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, *28*(5), 94–115. https://doi.org/10.1109/MSP.2011.941851

Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., … Y. Bengio (2015). EmoNets: Multimodal deep learning approaches for emotion recognition in video. *ArXiv:1503.01800 [Cs]*. http://arxiv.org/abs/1503.01800

Kim, H.-R., Kim, Y.-S., Kim, S. J., & Lee, I.-K. (2017). Building Emotional Machines: Recognizing Image Emotions through Deep Neural Networks. *ArXiv:1705.07543 [Cs]*. http://arxiv.org/abs/1705.07543

Kim, Y., Lee, H., & Provost, E. M. (2013). Deep learning for robust feature generation in audio-visual emotion recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3687–3691). Vancouver, BC, Canada, 2013. https://doi.org/10.1109/ICASSP.2013.6638346

Kollias, D., Marandianos, G., Raouzaiou, A., & Stafylopatis, A.-G. (2015). Interweaving deep learning and semantic techniques for emotion analysis in human-machine interaction. *10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (pp. 8766–8773). Trento, Italy, 2015. https://doi.org/10.1109/ICPR48806.2021.9413144

Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2020). Context based emotion recognition using EMOTIC dataset. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, *42*(11), 2755–2766. https://doi.org/10.1109/TPAMI.2019.2916866

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kumar, P., Jain, S., Raman, B., Roy, P. P., & Iwamura, M. (2021). End-to-end triplet loss based emotion embedding system for speech emotion recognition. *25th International Conference on Pattern Recognition (ICPR)* (pp. 8766–8773). Milan, Italy, 2021. https://doi.org/10.1109/ICPR48806.2021.9413144

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the open affective standardized image set (OASIS). *Behavior Research Methods*, *49*(2), 457–470. https://doi.org/10.3758/s13428-016-0715-3

Landowska, A. (2019). Uncertainty in emotion recognition. *Journal of Information Communication & Ethics in Society*, *17*(3), 273–291. https://doi.org/10.1108/JICES-03-2019-0034

Lang, P. (1999). *International Affective Picture System (IAPS): Technical manual and affective ratings*. Birmingham, USA: University of Alabama-Birmingham.

Li, H., & Xu, H. (2020). Deep reinforcement learning for robust emotional classification in facial expression recognition. *Knowledge-Based Systems*, *204*, 106172. https://doi.org/10.1016/j.knosys.2020.106172

Lim, N. (2016). Cultural differences in emotion: Differences in emotional arousal level between the east and the west. *Integrative Medicine Research*, *5*(2), 105–109. https://doi.org/10.1016/j.imr.2016.03.004

Liu, D., Jiang, Y., Pei, M., & Liu, S. (2018). Emotional image color transfer via deep learning. *Pattern Recognition Letters*, *110*, 16–22. https://doi.org/10.1016/j.patrec.2018.03.015

Liu, H., Sun, H., Li, M., & Iida, M. (2020). Application of color featuring and deep learning in maize plant detection. *Remote Sensing*, *12*(14), 2229. https://doi.org/10.3390/rs12142229

Liu, S., & Pei, M. (2018). Texture-aware emotional color transfer between images. *IEEE Access*, *6*, 31375–31386. https://doi.org/10.1109/ACCESS.2018.2844540

Lu, X., Lin, Z., Jin, H., Yang, J., & Wang, J. Z. (2014). RAPID: Rating pictorial aesthetics using deep learning. *22nd ACM international conference on Multimedia (MM '14)* (pp. 457–466). New York, NY, USA, 2014. https://doi.org/10.1145/2647868.2654927

Lu, X., Sawant, N., Newman, M. G., Adams, R. B., Wang, J. Z., & Li, J. (2016). Identifying emotions aroused from paintings. In G. Hua, & H. Jégou (Eds.), *Computer vision – ECCV 2016 workshops* (Vol. 9913, pp. 48–63). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46604-0_4.

Lu, X., Suryanarayan, P., Adams, R. B., Li, J., Newman, M. G., & Wang, J. Z. (2012). On shape and the computability of emotions. *20th ACM international conference on Multimedia (MM '12)* (pp. 229–238). Association for Computing Machinery, New York, NY, USA, 2012. https://doi.org/10.1145/2393347.2393384

Machajdik, J., & Hanbury, A. (2010). Affective image classification using features inspired by psychology and art theory. *18th ACM international conference on Multimedia (MM '10)* (pp. 83–92). Association for Computing Machinery, New York, NY, USA, 2010. https://doi.org/10.1145/1873951.1873965

Mao, Q., Zhu, Q., Rao, Q., Jia, H., & Luo, S. (2019). Learning hierarchical emotion context for continuous dimensional emotion recognition from video sequences. *Ieee Access*, *7*, 62894–62903. https://doi.org/10.1109/ACCESS.2019.2916211

Minsky, M. (1986). *The society of mind*. New York, USA: Simon and Schuster.

Mohammad, S. M., & Kiritchenko, S. (2018). Wikiart emotions: An annotated dataset of emotions evoked by art. *11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, 2018. http://saifmohammad.com/WebDocs/lrec2018-paper-art-emotion.pdf

Ortony, A. (2022). Are all "basic emotions" emotions? A problem for the (basic) emotions construct. *Perspectives on Psychological Science*, *17*(1), 41–61. https://doi.org/10.1177/1745691620985415

Ou, L.-C., Luo, M. R., Woodcock, A., & Wright, A. (2004a). A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research & Application*, *29*(3), 232–240. https://doi.org/10.1002/col.20010

Ou, L.-C., Luo, M. R., Woodcock, A., & Wright, A. (2004b). A study of colour emotion and colour preference. Part II: Colour emotions for two-colour combinations. *Color Research & Application*, *29*(4), 292–298. https://doi.org/10.1002/col.20024

Ou, L.-C., Luo, M. R., Woodcock, A., & Wright, A. (2004c). A study of colour emotion and colour preference. Part III: Colour preference modeling. *Color Research & Application*, *29*(5), 381–389. https://doi.org/10.1002/col.20047

Picard, R. W. (1995). *Affective Computing* (p. 16) [Technical Report No. 321]. M.I.T. 20 Ames St., Cambridge, MA 02139.

Poterek, Q., Herrault, P.-A., Skupinski, G., & Sheeren, D. (2020). Deep learning for automatic colorization of legacy grayscale aerial photographs. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 2899–2915. https://doi.org/10.1109/JSTARS.2020.2992082

Priya, D. T., & Udayan, J. D. (2020). Affective emotion classification using feature vector of image based on visual concepts. *The International Journal of Electrical Engineering & Education*, *002072092093683*, https://doi.org/10.1177/0020720920936834

Putkinen, V., Nazari-Farsani, S., Seppala, K., Karjalainen, T., Sun, L., Karlsson, H. K., … L. Nummenmaa (2021). Decoding music-evoked emotions in the auditory and motor Cortex. *Cerebral Cortex*, *31(5)*, 2549–2560. https://doi.org/10.1093/cercor/bhaa373

Ram, V., Schaposnik, L. P., Konstantinou, N., Volkan, E., Papadatou-Pastou, M., Manav, B., … C. Mohr (2020). Extrapolating continuous color emotions through deep learning. *Physical Review Research*, *2*(3), 033350. https://doi.org/10.1103/PhysRevResearch.2.033350

Rao, T., Xu, M., & Xu, D. (2018). Learning Multi-level Deep Representations for Image Emotion Classification. *ArXiv:1611.07145 [Cs]*. http://arxiv.org/abs/1611.07145

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 512–519). Columbus, OH, USA, 2014. https://doi.org/10.1109/CVPRW.2014.131

Renò, V., Losapio, G., Forenza, F., Politi, T., Stella, E., Fanizza, C., … R. Maglietta (2020). Combined color semantics and deep learning for the automatic detection of dolphin dorsal fins. *Electronics*, *9*(5), 758. https://doi.org/10.3390/electronics9050758

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … L. Fei-Fei (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A. (2017). Cross-cultural similarities and differences in affective processing and expression. In *Emotions and affect in human factors and human-computer interaction* (pp. 123–141). Elsevier. https://doi.org/10.1016/B978-0-12-801851-4.00004-5

Samanta, A., & Guha, T. (2021). Emotion sensing from head motion capture. *Ieee Sensors Journal*, *21*(4), 5035–5043. https://doi.org/10.1109/JSEN.2020.3033431

Sapinski, T., Kaminska, D., Pelikant, A., & Anbarjafari, G. (2019). Emotion recognition from skeletal movements. *Entropy*, *21*(7), 646. https://doi.org/10.3390/e21070646

Sartori, A., Yan, Y., Ozbal, G., Salah, A. A. A., Salah, A. A., & Sebe, N. (2015). *Looking at Mondrian's Victory Boogie-Woogie: What Do I Feel?* 7.

Shao, Z., Zhou, W., Cheng, Q., Diao, C., & Zhang, L. (2015). An effective hyperspectral image retrieval method using integrated spectral and textural features. *Sensor Review*, *35*(3), 274–281. https://doi.org/10.1108/SR-10-2014-0716

Shen, J., Zheng, J., & Wang, X. (2021). MMTrans-MT: A framework for multimodal emotion recognition using multitask learning. *13th International Conference on Advanced Computational Intelligence (ICACI)* (pp. 52–59). Wanzhou, China, 2021. https://doi.org/10.1109/ICACI52617.2021.9435906

Sowmyayani, S., & Rani, P. A. J. (2022). Salient object based visual sentiment analysis by combining deep features and handcrafted features. *Multimedia Tools and Applications*, *81*(6), 7941–7955. https://doi.org/10.1007/s11042-022-11982-5

Stamatopoulou, D., & Cupchik, G. C. (2017). The feeling of the form: Style as dynamic 'textured' expression. *Art and Perception*, *5*(3), 262–298. https://doi.org/10.1163/22134913-00002066

Sun, B., & Saenko, K. (2016). *Deep CORAL: Correlation Alignment for Deep Domain Adaptation*. https://doi.org/10.48550/ARXIV.1607.01719

Sundaram, V., Ahmed, S., Muqtadeer, S. A., & Reddy, R. R. (2021). Emotion analysis in text using TF-IDF. *11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 292–297). Noida, India, 2021. https://doi.org/10.1109/Confluence51648.2021.9377159

Tadi Bani, N., & Fekri-Ershad, S. (2019). Content-based image retrieval based on combination of texture and colour information extracted in spatial and frequency domains. *The Electronic Library*, *37*(4), 650–666. https://doi.org/10.1108/EL-03-2019-0067

Takada, A., Wang, X., & Yamasaki, T. (2021). Color-grayscale-pair image sentiment dataset and its application to sentiment-driven image color conversion. *2021 International Joint Workshop on Multimedia Artworks Analysis and Attractiveness Computing in Multimedia 2021 (MMArt-ACM '21)* (pp. 2–7). Association for Computing Machinery, New York, NY, USA, 2021. https://doi.org/10.1145/3463946.3469240

Tashu, T. M., Hajiyeva, S., & Horvath, T. (2021). Multimodal emotion recognition from art using sequential co-attention. *Journal of Imaging*, *7*(8), 157. https://doi.org/10.3390/jimaging7080157

Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., & Tesconi, M. (2017). Cross-Media learning for image sentiment analysis in the wild, *IEEE International Conference on Computer Vision Workshops (ICCVW)* (pp. 308–317). Venice, Italy, 2017. https://doi.org/10.1109/ICCVW.2017.45

Van, L. T., Nguyen, Q. H., & Le, T. D. T. (2022). Emotion recognition with capsule neural network. *Computer Systems Science and Engineering*, *41*(3), 1083–1098. https://doi.org/10.32604/csse.2022.021635

Vedaldi, A., & Soatto, S. (2008). Quick shift and kernel methods for mode seeking. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer vision – ECCV 2008* (Vol. 5305, pp. 705–718). Belin: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-88693-8_52.

Wang, M., & Deng, W. (2018). Deep Visual Domain Adaptation: A Survey. *ArXiv:1802.03601 [Cs]*. http://arxiv.org/abs/1802.03601

Wang, S., Han, K., & Jin, J. (2019). Review of image low-level feature extraction methods for content-based image retrieval. *Sensor Review*, *39*(6), 783–809. https://doi.org/10.1108/SR-04-2019-0092

Wei, Q., Zhao, Y., Xu, Q., Li, L., He, J., Yu, L., & Sun, B. (2017). A new deep-learning framework for group emotion recognition. *19th ACM International Conference on Multimodal Interaction (ICMI '17)* (pp. 587–592). Association for Computing Machinery, New York, NY, USA, 2017. https://doi.org/10.1145/3136755.3143014

Westlake, N., Cai, H., & Hall, P. (2016). Detecting people in artwork with CNNs. In G. Hua, & H. Jégou (Eds.), *Computer vision – ECCV 2016 workshops* (Vol. 9913, pp. 825–841). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46604-0_57.

Wolff, J. (2005). The meanings of minimalism. *Contexts*, *4*(1), 65–71. https://doi.org/10.1525/ctx.2005.4.1.65

Xu, L., Xu, W., & Zhang, W. (2021). Multi-dimensional music emotion recognition incorporating convolutional neural networks and Plutchik's emotion wheel completed research. *27th Annual Americas Conference on Information Systems (AMCIS 2021)*, Atlanta, USA, 2021.

Xu, M., Xu, C., He, X., Jin, J. S., Luo, S., & Rui, Y. (2013). Hierarchical affective content analysis in arousal and valence dimensions. *Signal Processing*, *93*(8), 2140–2150. https://doi.org/10.1016/j.sigpro.2012.06.026

Yanulevskaya, V., Uijlings, J., Bruni, E., Sartori, A., Zamboni, E., Bacci, F., … N. Sebe (2012). In the eye of the beholder: Employing statistical analysis and eye tracking for analyzing abstract paintings. *Proceedings of the 20th ACM International Conference on Multimedia* (pp. 349–358). https://doi.org/10.1145/2393347.2393399

Yanulevskaya, V., van Gemert, J. C., Roth, K., Herbold, A. K., Sebe, N., & Geusebroek, J. M. (2008). Emotional valence categorization using holistic image features. *15th IEEE International Conference on Image Processing* (pp. 101–104). San Diego, CA, USA, 2008. https://doi.org/10.1109/ICIP.2008.4711701.

Yao, X., Zhao, S., Lai, Y.-K., She, D., Liang, J., & Yang, J. (2020). APSE: Attention-aware polarity-sensitive embedding for emotion-based image retrieval. *IEEE Transactions on Multimedia*, *23*, 4469–4482. https://doi.org/10.1109/TMM.2020.3042664

Zhang, C., & Xue, L. (2021a). Autoencoder with emotion embedding for speech emotion recognition. *Ieee Access*, *9*, 51231–51241. https://doi.org/10.1109/ACCESS.2021.3069818

Zhang, C., & Xue, L.Two-stream Emotion-embedded Autoencoder for Speech Emotion Recognition. IEEE International IOT, Electronics and Mechatronics Conference *(IEMTRONICS)* (pp. 1–6). Toronto, ON, Canada, 2021. https://doi.org/10.1109/IEMTRONICS52119.2021.9422602

Zhang, H., Augilius, E., Honkela, T., Laaksonen, J., Gamper, H., & Alene, H. (2011). Analyzing emotional semantics of abstract art using low-level image features. In J. Gama, E. Bradley, & J. Hollmen (Eds.), *Advances in intelligent data analysis X: Ida 2011* (Vol. 7014). Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-24800-9_38

Zhang, J., Liu, X., Chen, M., Ye, Q., & Wang, Z. (2022). Image sentiment classification via multi-level sentiment region correlation analysis. *Neurocomputing*, *469*, 221–233. https://doi.org/10.1016/j.neucom.2021.10.062

Zhang, W., He, X., & Lu, W. (2020). Exploring discriminative representations for image emotion recognition with CNNs. *Ieee Transactions on Multimedia*, *22*(2), 515–523. https://doi.org/10.1109/TMM.2019.2928998

Zhao, S., Gao, Y., Jiang, X., Yao, H., Chua, T.-S., & Sun, X. (2014). Exploring principles-of-art features for image emotion recognition. *22nd ACM international conference on Multimedia (MM '14)* (pp. 47–56). Association for Computing Machinery, New York, NY, USA, 2014. https://doi.org/10.1145/2647868.2654930

Zhao, S., Huang, Q., Tang, Y., Yao, X., Yang, J., Ding, G., & Schuller, B. W. (2022). Computational emotion analysis from images: Recent advances and future directions. In B. Ionescu, W. A. Bainbridge, & N. Murray (Eds.), *Human perception of visual information* (pp. 85–113). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-81465-6_4.

Zhao, S., Yao, H., Gao, Y., Ji, R., Xie, W., Jiang, X., & Chua, T.-S. (2016). Predicting personalized emotion perceptions of social images. *24th ACM international conference on Multimedia (MM '16)* (pp. 1385–1394). Association for Computing Machinery, New York, NY, USA, 2016. https://doi.org/10.1145/2964284.2964289

Zhou, Y., Liang, X., Gu, Y., Yin, Y., & Yao, L. (2022). Multi-classifier interactive learning for ambiguous speech emotion recognition. *IEEE-ACM Transactions on Audio Speech and Language Processing*, *30*, 695–705. https://doi.org/10.1109/TASLP.2022.3145287

Zhu, Y., Zhuang, F., Wang, J., Chen, J., Shi, Z., Wu, W., & He, Q. (2022). *Multi-Representation Adaptation Network for Cross-domain Image Classification*. https://doi.org/10.48550/ARXIV.2201.01002.

Author Biographies

César González-Martín, PhD in Art, is full time professor and researcher in art at the University of Cordoba (Spain). He is expert in art research, contemporary art practices, design, and contemporary crafts. He has participated in several research projects like "visual arts, talent management and cultural marketing (ARTAPP)", and "MAKER ART: proposal for the digital transformation of the cultural industry related to handicrafts", and European projects like "Reuse Reduce Recycle AI-based platform for automated and scalable Maker culture in Circular economy (RRREMAKER)", "loW Altitude Remote sensing for the Monitoring of the state of Cultural hEritage Sites: building an inTegrated model for maintenance (WARMEST) or "GLobal cOntemporary art market: the intrinsiC and sociologicAL components of FINancial and artistic valuE of ARTtworks (GLOCALFINEART)".

Miguel Carrasco received his PhD in Informatics with highest-honors from Pierre et Marie Curie University-Paris 6 at the Institut des Systèmes Intelligents et de Robotique (ISIR) in 2010, and a PhD in Engineering Sciences in Computer Science from Pontificia Universidad Católica de Chile at the Department of Computer Science under a cooperation agreement on Jointly Supervised PhD, from which he was awarded with a scholarship from the Collège Doctoral Franco-Chilien in 2007. Currently, he is teaching at the Faculty of Engineering and Science from Universidad Adolfo Ibáñez (UAI). His research interest includes image processing, automatic visual inspection, computer vision, and pattern recognition.

Thomas Gustavo Wachter Wielandt is currently an artificial intelligence master's student at Utrecht University in The Netherlands. He received his bachelor's in computer engineering at the Adolfo Ibañez University in Chile, where he focused mainly on machine learning and computer vision. Now he is currently interested in the intersection of AI and philosophy, namely philosophy of mind, the epistemology of large language models, and the societal and ethical implications of technology in general.