RESEARCH ARTICLE

Predicting Emotion From Color Present in Images and Video Excerpts by Machine Learning

ALEKSANDRA WĘDOŁOWSKA[1], DAWID WEBER[2], AND
BOŻENA KOSTEK [ID][3], (Senior Member, IEEE)
[1]Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, 80-233 Gdańsk, Poland
[2]Department of Multimedia Systems, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, 80-233 Gdańsk, Poland
[3]Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications, and Informatics, Gdańsk University of Technology, 80-233 Gdańsk, Poland

Corresponding author: Bożena Kostek (bokostek@audioakustyka.org)

ABSTRACT This work aims at predicting emotion based on the colors present in images and video excerpts using a machine-learning approach. The purpose of this paper is threefold: (a) to develop a machine-learning algorithm that classifies emotions based on the color present in an image, (b) to select the best-performing algorithm from the first phase and apply it to film excerpt emotion analysis based on colors, (c) to design an online survey to check the accuracy of the annotations of the collected film data. In the first, three approaches to color extraction are tested, namely clustering colors into a palette of predefined colors, assigning colors to the RYB (Red, Yellow, Blue) model, and extracting a histogram of colors present in an image. This is based on image datasets containing color and emotion annotations. Classification is conducted using several algorithms, both deep learning and baseline artificial intelligence algorithms. The obtained results, under different configurations of parameters and training sets, are then presented. In the second part, the best-performing algorithm from the first phase is applied to film excerpt emotion analysis based on colors. This is followed by the third part, which is an online survey created to check the accuracy of the algorithm's annotations to the collected film data by the questionnaire respondents. Further, a discussion of the results achieved is presented. Conclusions contain a summary of the results and further direction for improving the performance of the created algorithm.

INDEX TERMS Color, emotion, machine learning, qualitative research, survey appraisal.

I. INTRODUCTION

Many factors affect human emotions when viewing images or watching films, one of which is the colors present in a scene. The work aims to develop a framework for predicting emotions based on colors in images, photography artwork, and film excerpts employing machine learning. Even though emotion and cognition, as well as predicting emotions by machine learning, may be treated as separate areas, when they interact, all these domains can benefit. Moreover, cognition understood as memory, attention, language, problem-solving, and reasoning processes, has all these factors in common with state-of-the-art machine learning algorithms [1].

There are a variety of definitions of emotion, as its derivation depends on the application field. The lack of an explicit description of emotion causes great difficulty in studying, modeling, and quantifying it [2]. Levenson, in ''Emotion, biology, and culture,'' defines emotion as a short-term psychological and physiological phenomenon adapting to a changing environment [3]. The title of this work is justified by showing the influence of biology and culture on the emotional system. The role of various factors in understanding emotion is also analyzed. They show that evolutionary, anthropology, social, biology, physiology, appraisal, biocultural, and even lexical and linguistic theories may be necessary for people's understanding of emotion.

Overall, human emotions are influenced by many aspects connected to the human senses; the primary five, i.e., sight, smell, touch, taste, and hearing, but also those treated as secondary, such as, e.g., nociception, a unique sensory system that refers to pain. In addition, the influence of color, music, environment, weather, etc., is also apparent. Generally,

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Oliva [ID].

the cause of emotional arousal is easier to recognize than what measures should be employed to label it. Even more challenging is how to evaluate emotional states. This also concerns affective analysis, in which sensory signals should be measured and considered in the feedback loop regarding the activity performed [4]. Moreover, another challenge is to map user experience based on emotion recognition and then adapt it to the current emotional state in an affective feedback loop [5].

Artists try to arouse emotions in the viewer by using a mixture of many elements, color being one of them. After popularizing color films, screenwriters began using the medium to give viewers additional information about the story they wanted to tell [6]. Recognizing what emotions a scene in a film should evoke is a complicated process and, again, depends on many factors, including color. The color scheme of an environment refers significantly to the mood of an entire scene and, therefore, the emotions it is trying to evoke.

The effect of colors on emotions is a topic that many researchers have explored, so it is worthwhile to look at recent solutions that aim to identify and predict emotions based on colors. These works concern abstract images [7], Iran-Islamic paintings [8], collecting images for exploration [9], creating an emotional movie database, psychophysiological assessing evoked emotions [10], employing, and analyzing human emotion changes quantitatively over time [11]. One interesting work is related to classifying emotions based on one feature only, i.e., a histogram with the Optimized RUSboosted Tree (CORT). The robustness of the CORT method was validated on three image datasets, and the method proved to be better than state-of-the-art prediction results [12].

Baveye et al.'s work refers to affective computing based on video content analysis [13]. Their work provides an overview of video databases used in movie content analysis, showing an accuracy of emotion recognition in the range of approx. 41% to 64%, depending on the dataset, number of annotators, emotion model, and method used. It also presents examples of measures related to, e.g., violence detection. Values of accuracy or agreement may be as low as 52.63% or as high as 85.5%. It should, however, be noted that a direct comparison between studies is difficult, as they differ in methods, emotion models, datasets, annotation quality, etc.

Wei et al. proposed a key frame-based extraction algorithm based on affective saliency estimation [14]. Key frames were extracted to avoid emotion-independent frames biasing the recognition results. It should be noted that this phase included object and semantic features from pre-trained ResNet-101 [15] and DeepSentiBank [16] models and combined them [14]. The emotion recognition was based on Support Vector Machines (SVM), Random Forests (RF), and Convolutional Neural Networks (CNN), as well as a hybrid fusion method proposed by the authors. For classification, six labels of emotions (anger, disgust, fear, joy, sadness, and surprise) were used. Experiments were performed on two datasets (Ekman-6 and VideoEmotion-8) with an average recognition for the fusion methods equaling 59.51% and 52.85%, respectively [14].

Another approach to Video Affective Content Analysis (VACA), i.e., automatic prediction of the emotional response of the viewers to a video, was proposed by Dudzik et al. [17]. They collected a crowd-sourced-based dataset called Mementos, which contains detailed information about viewers' responses to a series of music videos. This dataset includes descriptions of emotions of the watched videos and free-text descriptions of memories that were triggered while watching them. Two models were explored, i.e., the Early Fusion model employing feature vector containing memory-, visual-, and audio-based parameters fed to a Support Vector Regressor (SVR), and the Late Fusion approach with two separate SVRs for predictions using audio and visual features along with a Random Forest Regressor employing memory descriptions. The first experiment performed by these researchers was based on training the two proposed models with only memory descriptions. In contrast, the second experiment used different sources for training and test the models. For each experiment, the final predictions were for pleasure, arousal, and dominance. For that purpose, the R^2 (variance) score was calculated. It was found that an average value of R^2 for audiovisual tracks combined with memories has a greater value for each label (pleasure, arousal, and dominance). In conclusion, these authors showed that analyzing viewers' responses provided additional context for personalizing prediction in VACA, independent of the fusion strategy [17].

Chua et al. presented another approach to emotion prediction from music and video [18]. For this purpose, the researchers prepared a new dataset called MuVi, which is available in the public domain. The MuVi dataset consists of music videos in three modalities: audiovisual (original music videos), music-only, and visual (muted video only). The annotations were taken from the extended version of the Geneva Emotional Music Scale (GEMS-25 and GEMS-28). Moreover, there was additional annotation for the dataset in the arousal and valence space. Also, for GEMS annotations and arousal/valence, a contingency Chi-square statistical test was performed to check whether the frequency distribution of the emotion labels is significantly different for media stimuli presented in different modalities. For each media item, a set of audio and video features windowed over 500 ms without overlapping was extracted. The OpenSmile library extracted audio features, including MFCCs, pitch, spectral, zero-crossing rate, etc. [19]. For video features, six types of visual features were used, e.g., color, lighting key, facial expression, scene, and objects. The last three features were extracted using deep neural networks: a pre-trained model of VGG19, ResNet50, and Yolov5. For classification, multimodal architectures based on a unimodal architecture were used; all models were based on LSTM (Long Short-Term Memory) networks. Three stages of neural network models were proposed: early fusion based on two separate inputs

of audio and visual features, late fusion with a fully connected layer for prediction, and finally, PAIR (Predictive models Augmented with Isolated modality Ratings) with a pre-trained LSTM model and fine-tuning the weights of the multimodal audiovisual modality. The RMSE (Root Mean Squared Error) and CCC (Concordance Correlation Coefficient), i.e., their mean and standard deviation values, were calculated. For arousal, the best score of CCC was achieved only for the music modality, i.e., 0.4062 ± 0.32 and for valence, the best score of CCC was achieved on the visual modality, i.e., 0.3115 ± 0.32 [18].

Muszynski et al. proposed emotion data description on the (-1 to 1) range in the Valence/Arousal space [20]. For this purpose, the Continuous LIRIS-ACCEDE dataset, which contains 442 minutes of film fragments, was used. The machine learning models used audio-video parameters, high-level lexical parameters of text, and emotion annotations from film excerpts. Audio-video parameters were extracted with the openSMILE tool, and then the ReliefF algorithm was used to reduce the openSMILE features to 100 audio and 100 video parameters. 63 text features were prepared with Natural Language Toolkit. In addition, six features related to aesthetics were extracted. The LSTM network was used as the primary deep-learning model for classification. Then the LSTM accuracy was compared to the SVM and DBN (Deep Belief Network) models. The MSE values were calculated for each model, including for audio and video. The LSTM network achieved 0.045 for A-MES and 0.057 for V-MSEs [20].

Another proposal for using audio and video was described in Wang et al.'s work [21]. 31 audio and three visual features based on the MediaEval2017 dataset were used, described by arousal and valence values in the [-1 to 1] range. For classification, the Multimodal Regression Bayesian Network (MMRBN) was used and then compared to other methods. First, single RBN models were trained using audiovisual features. This was followed by stacking them into the form of two networks. The results of the MMRBN were compared to other methods used in the MediaEval task within the same conditions. The proposed MMDRBN showed relatively good performance on three datasets. Specifically, on the MediaEval2015 dataset, the accuracy was 44.26% for valence and 64.30% for arousal. On the MediaEval2016 dataset, the MSE (Mean Squared Error) was 0.332, and PCC (Pearson Correlation Coefficient) was 0.387 in the valence space. For arousal, the MMDRBN achieved 0.766 for MSE and 0.416 for PCC. The last test for MediaEval2017 returned an MSE of 0.110, and a PCC of 0.351 for valence. For arousal, MMDRBN achieved 0.103 for MSE and 0.315 for PCC. The results on arousal were higher than that of valence for the three datasets in comparison to other algorithms and works [21].

The present work aims to develop and apply machine learning algorithms that predict emotions based on the colors present in images and film excerpts. The structure of this work is threefold, as presented in Figure 1. The starting point is to test the accuracy of the algorithms in extracting color from images. Three approaches are tested, namely clustering colors into a palette of predefined colors, assigning colors to the RYB (Red, Yellow, Blue) model, and extracting a histogram of colors present in an image. For this purpose, datasets containing color and emotion annotations are employed. Four convolutional neural networks (CNNs), namely VGG16, VGG19, ResNet50, and DenseNet201, three types of baseline classifiers (logistic regression, decision tree, and random forest), as well as a recurrent neural network with an LSTM (Long Short-Term Memory) layer (keras.applications), were adopted for this study. The performance of the algorithms used is examined. The best-performing algorithm is then employed for predicting emotion in film excerpts according to the color extraction schemes tested on images.

Further, a survey is constructed to see to what degree the algorithm prediction agrees with the respondents' annotations. This is followed by a discussion of the results obtained and overall conclusions from the conducted experiments. Finally, possibilities for developing the framework proposed toward its more robust applicability are outlined.

II. EMOTIONS VS COLOR ANALYSIS
A. COLOR PERCEPTION
Color perception is possible through electromagnetic waves that are reflected, transmitted, or emitted by an object. Sensitivity to the length of these wavelengths makes it possible to distinguish colors. Human vision can see only a narrow range of electromagnetic wavelengths. Humans possess trichromatic vision, meaning that light is perceived through three types of cones located in the retina [22], [23], [24]. Each type of cone is made up of cells sensitive to different wavelengths of light. This results in the human recognition of three primary colors − red, green, and blue. Newton's theory of light states that white light in the visible spectrum is a combination of all other colors. White light can be split into seven primary colors: red, orange, yellow, green, blue, indigo, and violet [22], using a prism. Combining these colors in varying proportions enables humans to discern other colors [25]. It is worth noting, however, that 12% of women can perceive colors with an additional cone (tetrachromacy) [26].

Color vision can be described in three perceptual dimensions − hue, saturation, and brightness. Hue is the most distinctive feature as a healthy human eye can distinguish more than 200 colors. Colors that have hue are called chromatic colors. While those that do not are called achromatic. Examples of achromatic colors are white, gray, and the color created when a black surface is illuminated with white light. Chromatic and achromatic colors both have dark and light colors [27]. Of the existing colors, red, green, blue, and yellow are not mixtures of others [28]. The remaining colors (called binary) are combinations of the following pairs:

- green-yellow and yellow-green colors are mixtures of green and yellow,
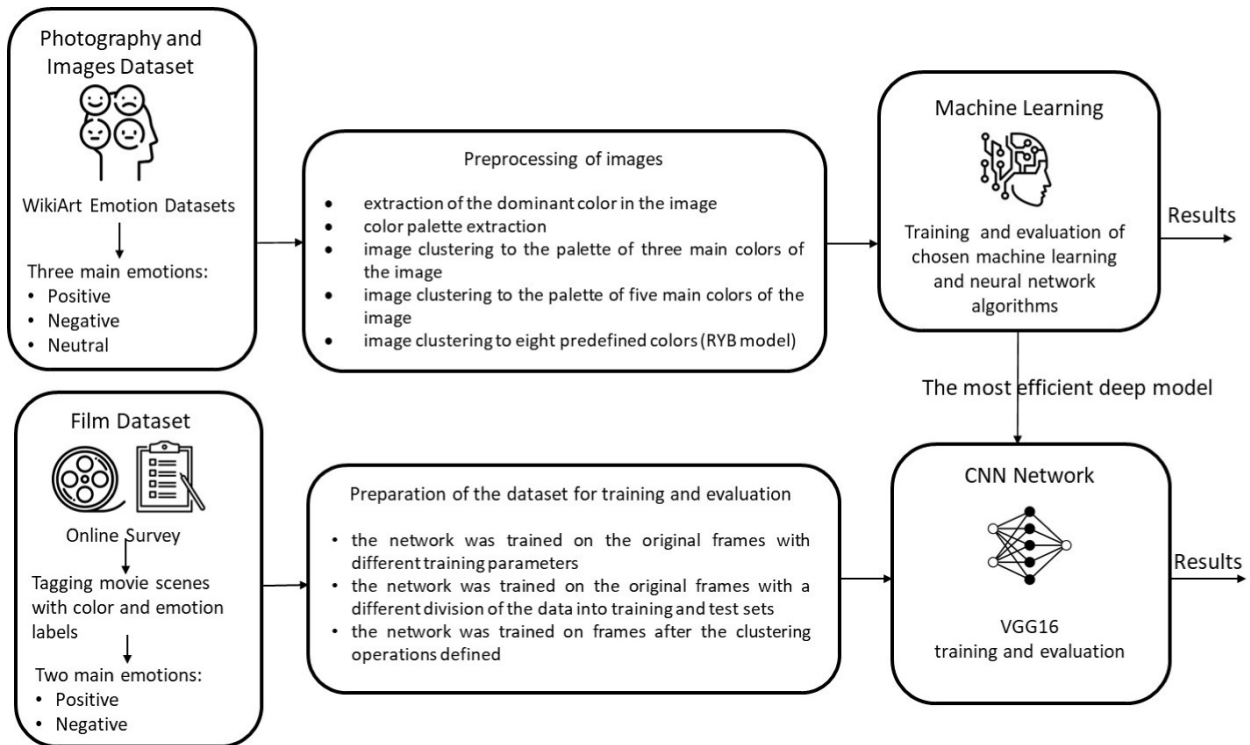- orange is a mixture of yellow and red,

FIGURE 1. Schema of the experimental design.

- colors such as magenta or purple are the result of a combination of red and blue,
- green-blue and blue-green colors are mixtures of green and blue.

In contrast to hue, saturation – the perceived purity of color, and brightness – the perceived intensity, as opposed to colorfulness, i.e., the absolute color intensity of a light stimulus, are relatively neglected in color labeling [28]. They are also less frequent in color-emotion-based analysis [29]. In this context, a question arises concerning what features of color are salient and should be used when correlating them with emotion.

B. ASSIGNING EMOTION TO COLOR

In everyday life, people very often unconsciously connect their emotions with colors. Examples include sayings such as ''black thoughts,'' ''look at/see something through rose-colored/tinted glasses,'' and ''act like a red rag to a bull.'' From these, one can draw simple conclusions about the meanings of colors: black -> sad, pink -> happy, red -> angry. Another use of colors to show emotion is emoticons – sending a message with a red icon tells of aggressive feelings, and in most cases, a green emoticon tells of disgust.

Takei and Imaizumi examined how the color-emotion association affects perceptual evaluation [30]. Moreover, they considered temporal proximity between color and facial stimuli interacting with these effects. They found that when the background color (yellow or red) corresponded to the given emotion (happiness or anger), it facilitated recognition of facial expressions. However, this effect occurred only when the background color was presented simultaneously with facial stimuli. Another color-emotion association described in the experiments was blue/gray -> sadness; however, it was evaluated as weak [30].

It should, however, be remembered that colors do not affect everyone in the same way. The personality and mental state of the subject can affect emotions. In an article by Bleicher, the effects of red and non-red colors in the Holtzman Inkblot Technique (HIT) were studied on five groups of people, namely: healthy, those suffering from neurotic disorders, borderline personality disorder, and acute and chronic schizophrenics [31]. The HIT technique involves testing patients' associations to ink blots [32]. Sets of ink blot cards are selected in HIT, taking into account whether the blot chosen is capable of arousing the subject's perception of detail, space, color, or shadows. The rule of thumb during HIT is to limit the patient to one response for each card. It was noted that people with schizophrenia distinguished themselves from the other groups when comparing responses for colored and black-and-white cards. In most groups, color on the cards resulted in different reactions than for cards without color. People with schizophrenia did not change their answers about emotions when color was introduced into the cards.

A person's mental state can affect the perception of colors and the emotions that come from them, but it is not the

only factor. The culture the person surveyed belongs to plays a massive role in color-emotion associations. For example, in European culture, black is associated with death and mourning, but nowadays, it is also regarded as a color associated with elegance. In China, black is the color of happiness and youth, while in ancient Egypt and India, it was the color of life. In Japan, black signifies mastery, hence the origin of the black belt in martial arts that speaks of the highest rank, while the white belt represents a beginner. Such differences are also associated with red, white, and even blue and violet.

Park and Guerin performed a series of experiments to prove the difference in color perception across cultures. Most of the results of these analyses supported the hypothesis that different cultures perceive colors differently, with the descriptive words of one palette differing by 93% across cultures [33].

Luscher's Psychological Color Test is also visible in the literature on color-emotion perception [34]. The first step in the Lüscher test is to prepare eight color cards. These cards are dark blue, blue-green, orange-red, bright yellow, purple, brown, black, and gray. Each card has a number written on the back. Each card has a meaning, which is analyzed later in the test. Colors are divided into primary (e.g., dark blue) and secondary (e.g., violet). It is necessary to choose a color that, without associations, evokes a positive reaction in the respondent. The test is repeated several times, and the results are grouped to assign the following meanings: favorite colors, the current state of the respondent, character traits not active at this time in life, and colors disliked by the respondent. This test seems complicated; it is, however, frequently used in emotion-color exploration; e.g., Ranjgar employed the Lüscher test to classify eight emotions that are evoked in the observer by Iranian-Islamic abstract paintings [8]. However, several modifications were introduced in the test preparation. The colors in the palette were arranged according to their frequency of occurrence in the image. Since background colors are not insignificant, they can affect the perception of the overall picture. For this reason, Ranjgar approached the problem using two perspectives, looking at the whole image and a portion of it [8].

In another study, the use of fuzzy color histograms to study the correlation between the color in an abstract painting and the emotions it evokes was proposed [7]. Ten colors were used to categorize the emotions evoked by an image. The colors selected were black, dark gray, red, brown, yellow, green, blue, cyan, magenta, and white. Eight emotion types were employed to describe the images — amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. An algorithm was created to study the relationship between emotion and color based on two main components — a color vector obtained from the survey data and an emotion vector. The emotion vector, a 10-color category histogram, built employing fuzzy logic, was used to obtain the color vector. Hibadullah confirmed a high correlation between the color black and the feeling of fear [7]. Also, cyan was highly correlated with contentment, yellow with amusement, and blue with disgust. Some colors (brown and green) showed

no significant correlation with any emotion from the list. In addition, some emotions were not strongly linked to any colors from the palette (e.g., rage and contentment). The remaining colors and emotions were linked to each other with a significant value of $p < 0.05$ or $p < 0.01$ [7].

Liu and Pei used a convolutional neural network and color transfer to colorize images based on user-defined emotions [9]. The algorithm has two paths. One possibility is to upload a reference image along with the target image. From the reference image, the main colors and texture features are extracted. Based on these features, emotion extraction then takes place. Another scenario is to skip the reference image and upload, along with the target image, the word associated with the emotion. Once the emotion is known (either resulting from the image or as annotated input), it is converted into color. This color is then used to change the color scheme of the target image so its color palette matches the emotion previously received. The authors decided not to use color histograms, as they only show the distribution of colors. Instead, they used harmonic models that demonstrate the hue distribution and can determine the relative relationship between colors. Considering the texture present in the image as an additional factor for analysis improved the results. They also showed that image segmentation plays a vital role in extracting the main colors. When a color is segmented, each image unit is treated as a whole, not resulting in scattered units [9].

To evaluate the performance of the created method, Liu and Pei designed a model that calculated the emotion evoked by the final image and checked the similarity of this emotion to the input value. Comparing the algorithm's performance with state-of-the-art methods showed much better correspondence between emotions and color when employing seven-color models. This is because for images with multiple colors, three- and five-color models may be insufficient. So, this is one of the important conclusions derived from this study [9].

III. METHOD
A. DATA EMPLOYED
In the work carried out, it was necessary to prepare a set of images and film excerpts to develop to perform color analysis and automatically assign them to the appropriate emotion classes. In addition, a survey was created to collect respondents' emotion label assignment to the film excerpts, aiming at discovering color-emotion associations.

B. IMAGES AND PHOTOGRAPHY SETS
The WikiArt emotions dataset is a subset of WikiArt [35] that consists of a collection of more than 150,000 images of artworks in 14 different styles, including modern art, Japanese art, Korean art, and Western Renaissance art. The database contains works by 195 artists, which can be divided into 168 categories (for example, manga, portrait, sculpture, or flower paintings). The emotion subset contains

4,105 artworks from the WikiArt database. These are primarily paintings from 22 categories (for example, realism or impressionism) and four styles − Renaissance art, post-Renaissance art, contemporary art, and custom art.

The annotations on the collection were made through crowdsourcing − a process involving a large group of people working on a specific task. Each artwork in the database was annotated by a minimum of 10 people. Annotations were made in three scenarios:

- presentation of only the image (without the title) and asking what emotions the image evokes in the viewer;
- presentation of only the title of the artwork (without the painting) and asking what emotions the title evokes in the viewer;
- presentation of the entire work of art, the painting with the title, and the question of what emotions the work – as a whole – evokes.

The annotations of each painting in the database include the following information: the style of the artwork depicted; the category into which the image falls; the name of the artist; the title of the artwork; the year the work originated; the average rating by the respondents; etc.

An additional dataset analyzed in our work contains images from a collection incorporating three datasets. These are the International Affective Picture System (IAPS), artistic photography searched by keywords in a photo-sharing forum, and a set of abstract images rated against emotions by a group of people. The collections label eight emotions: amusement, anger, awe, contentment, disgust, excitement, fear, and sadness [36]. This dataset was also used by Machajdik and Hanbury [37].

C. IMAGE DATA PREPARATION

Pre-processing was required to use the data collected. Operations performed on the images included cleaning data, cropping the image, extracting the dominant color and color palette from the image, clustering the image to the dominant or predefined main colors found in the image, and extracting color information using histograms.

The original annotations in the WikiArt emotions database had emotion ratings after seeing the image, presenting the title itself, and seeing the entire artwork (image and title). There are 20 emotions labeled in the WikiArt database.

The decision on the emotion model or the number of emotions contained in experiments typically constitutes the first step of experiments or while creating a survey [38], [39], [40], [41]. In our study, to build a classifier of satisfactory effectiveness, we propose to map the original emotion assignment into three baseline categories – positive, negative, and neutral, as shown in Table 1. Overall, 1,427 positive, 553 negative, and 1,196 neutral labels were obtained. The difference in the original number of images and the number of images used in the experiments is because 929 images had no tags in the image-only rating category.

TABLE 1. Assignment of emotions present in the WIKIART emotions database.

Original Annotation	New Label
Agreeableness	Positive
Gratitude	
Happiness	
Love	
Optimism	
Trust	
Fear	Negative
Anger	
Arrogance	
Disagreeableness	
Disgust	
Pessimism	
Regret	
Sadness	
Shame	
Shyness	
Anticipation	Neutral
Humility	
Surprise	
Neutral	

D. EXTRACTION OF DOMINANT COLORS

The ColorThief library was used to extract the dominant color and the palette of the dominant colors in the image. This library has two methods corresponding to [42]:

- getColor, which accepts a quality argument affecting the quality of the result. The value of this parameter can be in the range of 1–10, where one means the best quality, but it increases the image analysis time. The method returns a tuple with color values in the (R, G, B) format,
- getPalette, which takes a colorCount parameter in addition to the quality argument, specifying how many colors the returned palette should contain. The result of the method is a list of tuples in the same format as for the getColor function.

E. CLUSTERING INTO A PALLET OF IMAGE COLORS

Clustering, or cluster analysis, allows similar data points to be grouped into clusters. One of the efficient cluster analysis methods is the K-means algorithm, an unsupervised machine-learning method (keras.app). The algorithm requires a predefined number of clusters for it to create.

The K-means algorithm from the sklearn.cluster library was used to analyze the clusters in the original image and transform it into an image containing only a certain number of colors. The method takes many parameters, of which one of the most important is $n\_clusters$, specifying the number of clusters to be created. The values returned are associated with labels assigned to each pixel and the coordinates of the centers of the clusters.

An example of the results of the cluster analysis for the three- and five-image clusters is shown in Figure 2.

FIGURE 2. Clustering results (a) original image, (b) 5 clusters, and (c) 3 clusters using keras.app.

F. CLUSTERING INTO A PREDEFINED COLOR SPACE

The previous method describes the clustering of image pixels to the main colors present in the image. This approach transformed each image after clustering into one with a color palette unique to the given image. Another way of clustering is to analyze and assign image pixels to predefined clusters of colors. Clustering into eight colors was chosen according to the RYB color space often used in art. Therefore, the primary colors (red, yellow, and blue), secondary (orange, purple, and green), and black and white, as the ones that augment a spectrum of colors when mixed with the primary colors, were selected.

To perform such a cluster analysis, KNN (K-nearest neighbors), one of the most popular supervised learning methods, was employed [12]. The principle of the algorithm is to find the best match to the test data. The KNN method was implemented using the OpenCV library and the KNearest_create and *findNearest* methods. An example of the results of clustering into the main predefined color is shown in Figure 3.

FIGURE 3. Example of results of clustering to main predefined colors.

G. HISTOGRAMS

Histograms were used as one of the methods of color extraction. The first step in extracting image features was to cluster up to five clusters using the K-means method described earlier [43]. The information about these clusters was then sent to the process responsible for creating the histogram. This was implemented using NumPy's histogram function, whose parameters are the colors obtained by clustering and the number of rectangles of the histogram. Then the histogram was normalized by dividing it by the sum of all the components.

The created and normalized histogram was then passed to a function whose task is to extract five parameters of color information and store them in a CSV file. This information is as follows:

- average value of R (red color),
- average value of G (green color),
- average value of B (blue color),
- color distribution,
- variance of brightness,
- saturation variance.

H. FILM EXCERPTS

30-second-long excerpts of films collected at the Gdansk University of Technology annotated with five different emotions were also employed [44]. These are excerpts from 30 film productions (see Table 2). The annotations were based on surveys of color-emotion associations, the context of the film, and the literature description. The tagged emotions are happiness, fear, excitement, sadness, and aggression, as well as "none assigned." The last case refers to contradictions met while assigning emotion to the film.

TABLE 2. Film excerpts, along with emotion annotation, used in the study.

Happiness	Sadness	Fear	Excitement	Aggression	None assigned
Amelia	Corpse Bride	Maleficent	Avatar	Deadpool	Guardians of the Galaxy
Harry Potter and the Philosopher's Stone	Spirited Away	Ace Ventura	Gladiator	Ex Machina	Into the Wild
Her	Grand Budapest Hotel	Kill Bill vol. 2	Lost River	Inglorious Basterds	The Machinist
Joker	Phantom	Pirates of the Caribbean	Rebel Without a Cause	The Shining	X-Men
The Godfather					
The Lion King					
The Martian					
The Royal Tenenbaums					
A Very Long Engagement					
Titanic					

FIGURE 4. Example of thresholding operation on a frame of "The Godfather" film.

FIGURE 5. Survey use diagram.

I. FILM DATA PREPARATION

One of the operations required for film-based data was segmentation, which was performed to obtain single frames. Segmentation of the videos from MP4 format into JPG (Joint Photographic Experts Group) format frames was performed using the *VideoCapture* method from the OpenCV library [45].

Since some of the video excerpts contained black frames around the video image, image cropping was applied so the algorithm could remove unwanted areas and focus on the actual colors present in the scene. The image cropping was performed using the OpenCV library and the *cv.threshold* function.

The process of cropping images for each video excerpt was as follows:

- Uploading the selected grayscale frame of the video;
- Sending the uploaded image to the thresholding method;
- Returning the image cropped by the thresholding method to the dimensions of the created mask;
- Checking the correctness of the performed thresholding and possibly changing the parameters of the method.

If the thresholding for each frame in the analyzed excerpt is performed correctly, then it is performed as follows:

- uploading the frame,
- applying the thresholding function,
- saving the image after thresholding.

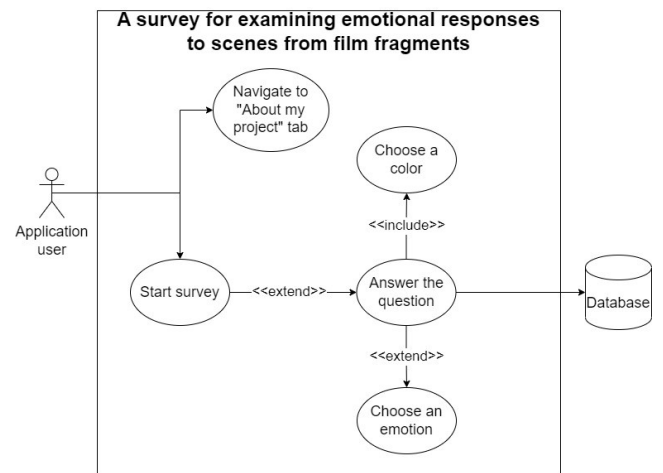An example of the thresholding operation working on a frame of "The Godfather" film is shown in Figure 4.

J. SURVEY – COLLECTING FEEDBACK FROM RESPONDENTS

The survey aimed to verify the correctness of the emotion annotation on the collected video. It notes the color the user chose as relevant to the emotion felt.

The survey was created in the form of a web page. The start screen includes instructions for using the questionnaire form, explains the purpose of collecting user responses, and illustrates a sample response to a question.

The presentation of the operation and handling of the survey is shown in Figure 5 in the form of a use case diagram executed in UML (Unified Modeling Language).

When entering the survey, the user sees the start page and can begin filling in the survey by clicking the appropriate button — a page with the first question then appears. It explains the purpose of data collection and a sample answer to the survey questions (see Figure 6). At any time, it is possible to return to the main page or an additional tab with more detailed information about the project.

As explained on the survey's start screen, after pressing "Start the questionnaire," the respondent is taken to a screen where 15 questions are displayed sequentially, each containing three scenes from a randomly selected movie, a place to choose an emotion, and one to choose a color.
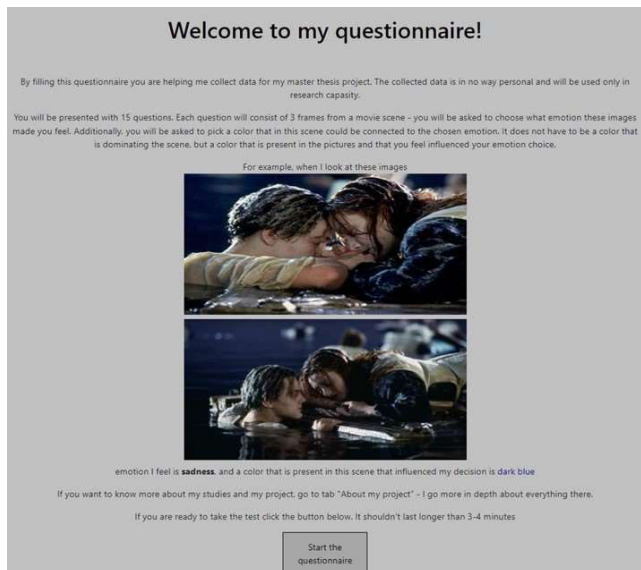
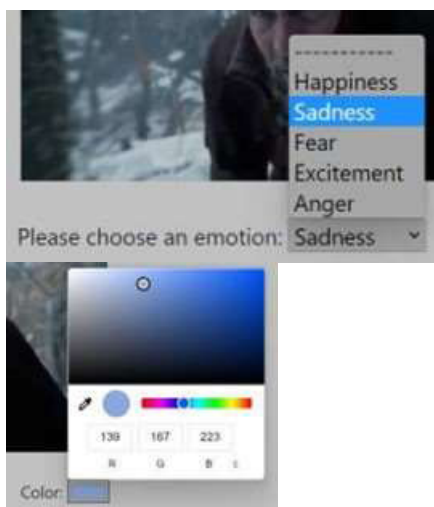**FIGURE 6.** Start screen of the survey.

FIGURE 7. Operation of the (a) emotion and (b) color checkboxes in the survey.

The emotion selection field is a mandatory single-choice field in the form of a disjoint list. The color selection has a default value set as black, and when the users click on the area, they can select any color stored in hex format. To choose a color, the respondent can use the cursor in the color map, type RGB values into the appropriate fields, or hover the cursor over the image and click to select the appropriate color directly from the image. The view of the selected areas in the survey is presented in Figure 7.

After answering all the questions, the respondent is taken to a thank you page where there is also a request for further distribution of the survey.

The database used in the survey is relatively small. It is a single table storing the response ID, the video's name, the emotion, and the color selected by the user.

IV. EXPERIMENTS
A. ALGORITHMS
Several experiments were conducted on available deep-learning models; however, they were modified according to the needs of the study. Their implementation was facilitated by the keras.applications library, which provides deep learning algorithms along with pre-trained weights. Also, three baseline algorithms were implemented for comparison of color-emotion assignment effectiveness. Finally, two types of convolutional neural networks and a deep model consisting of a recurrent neural network augmented with an LSTM (Long Short-Term Memory) layer completed the investigation.

- VGG16 (keras.app) – The VGG16 model with weights transferred was applied to classify three color classes. The weights used were those obtained from training on the *imagenet* set (Image database). Freezing the weights was performed by setting the trainable parameter to False for all the original model layers. This results in no training of these layers, which speeds up the training time and reduces the resources needed for training. However, four new layers were added to the model, i.e., Flatten, Dropout, and two Dense layers. The activation function of the last layer was softmax.
- VGG19 (keras.app) – The implementation of the VGG19 network proceeded similarly to the VGG16 model – four new layers were added, freezing the weights of the layers already contained in the model.
- ResNet50 (keras.app) – The ResNet50 model was also implemented with the transfer of weights but without freezing the training of the original model layers. Also, as in previous models, weights from the *imagenet*-based training were used. In the implementation of this model, six new layers were added – GlobalAveragePooling2D, two Dropout layers, and three Dense layers.
- DenseNet201 (keras.app) – In the implementation of the DenseNet201 model, the last original layers were removed in the model development; the initial implementation left the *include_top* parameter equal to True. For this model, four new layers were added, i.e., Flatten, Dropout, and two Dense layers.

In addition, three baseline classifiers were implemented: decision tree, random forest, and logistic regression.

- DecisionTreeClassifier [46] – To create a classifier, the DecisionTreeClassifier class was imported, and the fit and prediction methods were called. To evaluate the model, the accuracy_score method was imported from the sklearn.metrics library. The method used to find the best model parameters was GridSearchCV. The parameters searched were the maximum tree depth, the minimum number of samples needed to split a node, and the minimum number of samples required for a node to be a leaf.
- Random Forest classifier [46] – Random forest was implemented by the RandomForestClassifier class from the sklearn.ensemble library. As in the decision

tree implementation, GridSearchCV was used to improve the prediction results. The optimization consisted of finding the best hyperparameters for the number of estimators, the maximum depth of the tree, the minimum number of samples needed to split a node, and the minimum number of samples required for a node to be a leaf.

- Logistic regression − Logistic regression was implemented by the LogisticRegression class from the sklearn.linear_model library. To improve the performance of the model, the L1 (performing a linear transformation on the weights) and L2 (adding a squared cost function to your loss function) regularization methods were used [47].
- CNN (convolutional neural network) model − As part of the experiments on image datasets, 1-dimensional and 2-dimensional CNNs were created. The algorithms were implemented using the Keras library. Convolutional networks intended for 1-dimensional data with one pooling layer using the L2 regularization method and a network for image data containing two pooling layers were further explored.
- Recurrent model − It was also decided to use a recurrent approach to train the classification on the video excerpts. The scheme adopted in this case was feature extraction from the video frames, performed with the VGG16 model, and then the resulting information was used to train the model with the LSTM (Long Short-Term Memory) layer. The layers of the network with a recurrent layer are shown in Table 3. The activation functions for the Dense layers were the ReLu and Softmax functions.

TABLE 3. Architecture of a model with a recurrent layer.

Layer type	Output shape	Number of trainable parameters
LSTM	(None, 256)	787,456
Dense	(None, 1024)	263,168
Dropout	(None, 1024)	0
Dense	(None, 2)	2,050

V. RESULTS

A series of experiments were conducted on the algorithms described earlier. To obtain the highest accuracy, the training parameters were modified, followed by an analysis of the change in the score after transforming the training method.

The parameters modified during the experiments were:

- learning speed,
- batch size (batch),
- optimizer,
- use (or not) weight transfer.

To analyze the results and compare the classification performed by the algorithms, four basic metrics were used: accuracy, precision, sensitivity, and F1 score measure [48].

The experiments were divided into several stages based on the databases on which they were performed. The first was the WikiArt emotions database. As shown in Table 1, the emotion labels refer to three classes: positive (class no. 1), negative (class no. 2), and neutral (class no. 3). The database does not have an equal distribution of elements into classes – the category of positive emotions contains the most data, while the images that arouse negative emotions have the least. The exact proportions of the distribution of emotions into categories in the test set are given below:

- class 1: 83 images,
- class 2: 180 images,
- class 3: 214 images.

Data processing operations were performed on the WikiArt database, as described earlier:

- extraction of the dominant color in the image;
- color palette extraction;
- image clustering to a palette of the three main colors in the image;
- image clustering to a palette of the five main colors in the image;
- image clustering to eight predefined colors (RYB model).

Uploading the data to the input of the algorithms was carried out using the ImageDataGenerator method, which retrieves the data from the specified folder while performing the defined data augmentation. The parameters for augmentation used during the data download are:

- shear range,
- zoom range,
- horizontal image rotation,
- rotation range,
- width change range,
- height change range.

The results obtained were not overly satisfying. Nonetheless, here are the best results. For the VGG16 and VGG19 models, the average accuracy was approx. 64% for the three- and five-color classes, with an average F1 score of 41.46%. For the RYB-based predefined classes, VGG16 and VGG19 returned an average accuracy of approx. 61%. The results for DenseNet201 and ResNet50 were even less satisfying. Moreover, the other calculated measures were not satisfactory. However, when analyzing the results of precision and sensitivity, a relationship between them was observed − these measures have smaller values for classes with fewer objects.

When the classification process was repeated for the original images from WikiArt (without any processing), the average accuracy increased by about 3 percentage points, with the highest accuracy of 67.09 for VGG19. This may be due to the larger amount of data used for training and testing.

Deep learning network training was also conducted on the artistic photography dataset that contained images grouped

into one of three categories: positive emotions (class no. 1), negative emotions (class no. 2), and neutral (class no. 3). The distribution of data into classes for the test set was as follows:

- class 1: 61 images,
- class 2: 80 images,
- class 3: 22 images.

The results of training the neural networks on the original images of the database of photographs and abstract images were comparable to those obtained on the WikiArt emotions. The average accuracy for VGG16 and VGG19 was approx. 64%, whereas the DenseNet201 and ResNet50 models decreased to about 50%. When color scheme division was performed, the results decreased to about 60% for the VGG models.

Images from a database of photographs and abstract images and images from the WikiArt emotions database were used to train and test the baseline algorithms and CNN networks created with the Keras library. The results were even worse − the algorithms stopped learning at accuracy values of about 50%.

Finally, the most interesting part of the exploration was devoted to color classification on film excerpts. To analyze the colors and their change throughout the video, the algorithm on which the classification experiments for the video fragments were performed included a recursive layer.

Three experiments were conducted to test the effect of the training data on the algorithm's performance:

- the network was trained on the original frames with different training parameters;
- the network was trained on the original frames with a different division of the data into training and test sets;
- the network was trained on frames after the clustering operations defined earlier.

Due to the small amount of film excerpt data and insufficient variety of their emotion denotations, a simplified division of emotions into categories was used. Thus, emotions were divided into positive (happiness, excitement) and negative (fear, sadness, aggression), resulting in a binary division. Three training sets of video excerpts were created according to the class assignment, as shown in Table 4.

TABLE 4. Test collection configurations of film excerpts.

Set No.	New Label	Film Excerpts
1	Positive	Amelia, Harry Potter, Lost River
	Negative	The Witch, The Shining, X-men
2	Positive	Avatar, The Lion King
	Negative	Corpse Bride, Into the Wild, Phantom
3	Positive	Gladiator, The Godfather, Titanic
	Negative	Ex Machina, Kill Bill, Spirited Away

First, training and prediction were performed with different sets of hyperparameters. The results of the experiments with

three sets of hyperparameters are shown in Table 5. In addition, the classification results depending on the data division into training and test sets based on three collection sets are contained in Table 6. Highlighted values in Tables 5 and 6 refer to the best scores.

TABLE 5. RNN classification results on frames of video fragments for a variety of parameter settings.

Parameters	Class	Precision [%]	Sensitivity [%]	F1 score [%]	Accuracy [%]
Optimizer => SGD (Stochastic Gradient Decent) Learning rate = 0.00005	Positive	74	86	79	73.09
	Negative	71	54	61	
Optimizer => Adam learning rate = 0.00005	Positive	72	91	81	73.26
	Positive	77	46	57	
Optimizer => Adam learning rate = 0.00001	Positive	75	87	81	**74.62**
	Negative	74	56	63	

TABLE 6. Algorithm test results depending on the division of data into training and test sets for three collection set.

Parameters	Class	Precision [%]	Sensitivity [%]	F1 score [%]	Accuracy [%]
Optimizer => SGD Learning rate = 0.00005	Positive	74	86	79	**73.09**
	Negative	71	54	61	
Optimizer => Adam Learning rate = 0.00005	Positive	87	59	70	68.33
	Positive	54	85	66	
Optimizer => Adam Learning rate = 0.00001	Positive	77	55	65	61.55
	Negative	74	86	79	

The above experiments were conducted on the original film frames without performing clustering operations. Further, two types of clustering were performed on the images – to the main colors of the image (palettes of three and five colors) and clustering to eight colors (RYB model). The experiments were carried out on test set no. 1, and the results are shown in Table 7. Note that the highlighted value refers to the best score.

Data obtained from the color histogram were also used to analyze the film data. The mean values of the R, G, and B components, along with the color distribution, brightness variance, and saturation variance, were trained using the following algorithms: decision tree, random forest, logistic regression, and convolutional network (CNN). The prediction

TABLE 7. Test results depending on the clustering operation performed on the video frames.

Parameters	Class	Precision [%]	Sensitivity [%]	F1 score [%]	Accuracy [%]
Original	Positive	74	86	79	73.09
	Negative	71	54	61	
After clusterization to the palette of three colors	Positive	82	90	86	81.67
	Negative	81	69	75	
After clusterization to the palette of five colors	Positive	78	86	82	77.11
	Negative	74	64	69	
After clusterization to the RYB model-based colors	Positive	93	97	95	**93.58**
	Negative	95	89	92	

TABLE 8. Results of algorithm tests performed on information extracted from color histograms.

Algorithm	Version	Accuracy [%]
Decision tree	basic	52.83
	after optimization	60.48
Random forest	basic	58.46
	after optimization	66.11
Logistic regression	basic	50.42
	after optimization	53.88
CNN (regularization L2)	–	**73**

results on dataset no. 1 are shown in Table 8. The best score is indicated in bold font.

However, the model that achieves the best results on film data is RNN (augmented with LSTM) trained after clustering the images into RYB model-based colors and binary emotion division. With such a variant, the maximum accuracy reached the optimum and was higher than 95%. Plots of the accuracy values and the loss function during training for this model are shown in Figure 8.

Due to the high result obtained for the RNN model on the binary prediction of the film excerpts, training was also performed on splitting the data into the original five emotions. The results of the classification are shown in Table 9. As seen in Table 9, the results are not satisfactory apart from those obtained for happiness (indicated in bold font).

VI. DISCUSSION

The experiments were first performed on WikiArt emotions databases and a collection of photographs and abstract
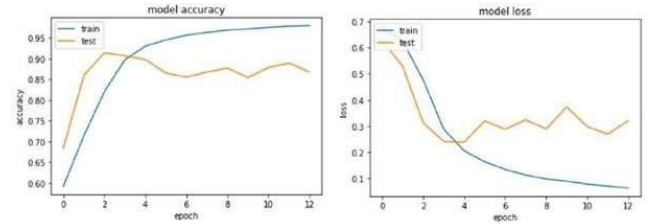

FIGURE 8. Accuracy and model loss for RNN on frames of video fragments after clustering operation into RYB model-based colors and binary division of emotion.

TABLE 9. Results of the RNN algorithm after training on the video data with a division into the five original emotions.

Class	Precision [%]	Sensitivity [%]	F1 score [%]	Accuracy [%]
Happiness	**81**	96	88	
Sadness	83	26	40	
Fear	50	53	52	67.93
Aggression	29	32	31	
Excitement	73	59	65	

images. Both datasets were divided into three categories of emotions: positive, neutral, and negative. A problem of class inequality was observed in both databases — the WikiArt emotions set contained a relatively small number of images evoking negative emotions. In contrast, the photography and abstract images database lacked images in the neutral category. The image databases were used to test both baseline and deep-learning algorithms. Based on the results obtained, it can be said that both VGG models outperformed the baseline and more complex models. The deep nature of DenseNet and ResNet makes them more prone to overlearning, especially with such a small amount of data as used in the experiments. However, it is difficult to compare the two VGG models – the results of their prediction are similar to each other.

A series of experiments were conducted to train algorithms on film data. Due to the problem of contradiction in data labeling, the film set was divided into two categories of emotions – positive and negative. The average accuracy value received from different dataset collections, contained in Table 4, is 67.66%, as shown in Table 6.

A further experiment was performed to reduce the amount of information in the image and focus only on the colors and their distribution in the scene. Clustering methods were used for this, and the results presented in Table 5 show that using each of the implemented clustering variants improves the accuracy of emotion prediction. Clustering the images into the RYB color model, or a palette of colors, improved the algorithm's ability to classify emotions. An average accuracy of over 81% was achieved when grouping images into a palette of three colors. Clustering into RYB model-based

colors, compared with the original video frames, raised the accuracy value by 20%, which translates into equally high precision, sensitivity, and F1 score values.

Further reducing the amount of information used in training was done using only the color values obtained from histogram-based processing. During the training, only information describing the colors and not their location in an image is given to the algorithm's input. As a result, baseline machine learning algorithms achieved results of less than 60% before the optimization operation. After optimization, two algorithms, i.e., decision tree and random forest, increased the prediction accuracy by about 8%. Using CNN for training on histogram color data resulted in better results. Compared to the machine learning algorithms, the increase in accuracy was more than 10% (as shown in Table 8). The final experiment consisting of training and testing with the division of film data into five emotions achieved satisfactory results for happiness and the least for aggression (Table 9).

A. SURVEY RESULTS

The database of films used had color and emotion designations. To check the accuracy of the initial (originally labeled) assignments, a survey was created with random questions about 15 of 30 films. Each question includes three scenes from a given film excerpt and a question about the emotions these scenes evoke in the respondent. In addition, there is a question about identifying the color whose presence in the scene supports the feeling of that emotion.

The questionnaire was completed by 45 people. Tables 10–14 compare the initial labels of the set of films with the tags obtained from the survey results. The tables are divided by the initial emotion assigned to the films.

TABLE 10. Comparison of initial labels with labels resulting from the survey when the initial emotion is fear.

Film title	Initial color assignment	Initial emotion assignment	Color as assigned by the survey respondents	Emotion as assigned by the survey respondents
Ace Ventura	Green	Fear	Red, black	Fear
Kill Bill vol.2	Green	Fear	Black	Fear
Maleficent	Green	Fear	Green, black	Fear
Pirates of the Caribbean	Green	Fear	Gray, blue	Fear, excitement
The Machinist	Green	Fear	Black, gray	Sadness
Into the Wild	Green	Fear	Green	Fear

In Tables 10–14, the emotion that received the highest number of votes in the survey conducted is given in the

TABLE 11. Comparison of initial labels with labels resulting from the survey when the initial emotion is excitement.

Film title	Initial color assignment	Initial emotion assignment	Color as assigned by the survey respondents	Emotion as assigned by the survey respondents
Avatar	Purple	Excitement	Dark blue	Excitement
Gladiator	Purple	Excitement	Dark blue	Fear, sadness
Lost River	Purple	Excitement	Purple	Excitement
Guardians of the Galaxy	Purple	Excitement	Violet	Fear
Avatar	Purple	Excitement	Dark blue	Excitement
Gladiator	Purple	Excitement	Dark blue	Fear, sadness

TABLE 12. Comparison of initial labels with labels resulting from the survey when the initial emotion is happiness.

Film title	Initial color assignment	Initial emotion assignment	Color as assigned by the survey respondents	Emotion as assigned by the survey respondents
Amelia	Orange	Happiness	Orange, yellow	Happiness
Harry Potter and the Philosopher's Stone	Orange	Happiness	Orange	Happiness
Her	Orange	Happiness	Beige	Happiness
Joker	Orange	Happiness	Orange, black	Excitement, happiness
The Godfather	Orange	Happiness	Pink	Happiness
The Lion King	Orange	Happiness	Blue, yellow	Happiness
The Martian	Orange	Happiness	Brown	Fear, excitement
The Royal Tenenbaums (excerpt no. 1)	Orange	Happiness	Orange	Happiness
The Royal Tenenbaums (excerpt no. 2)	Yellow	Happiness	Gray	Excitement
A Very Long Engagement	Orange	Happiness	Green, yellow	Happiness
Titanic	Orange	Happiness	Orange	Happiness

columns with the survey results. In cases where two emotions received the same number of votes, both were entered in the table. For example, regarding the "Joker" movie, seven respondents voted for the emotion "Excitement," and the other seven people for "Happiness."

TABLE 13. Comparison of initial labels with labels resulting from the survey when the initial emotion is sadness.

Film title	Initial color assignment	Initial emotion assignment	Color as assigned by the survey respondents	Emotion as assigned by the survey respondents
Corpse Bride	Blue	Sadness	Blue, dark blue	Sadness
Spirited Away	Blue	Sadness	Lilac, violet	Sadness
Grand Budapest Hotel	Blue	Sadness	Gray	Fear
Phantom	Blue	Sadness	Gray	Fear
X-Men	Blue	Sadness	Blue	Fear
Corpse Bride	Blue	Sadness	Blue, dark blue	Sadness

TABLE 14. Comparison of initial labels with labels resulting from the survey when the initial emotion is anger.

Film title	Initial color assignment	Initial emotion assignment	Color as assigned by the survey respondents	Emotion as assigned by the survey respondents
Deadpool	Red	Anger	Gray, blue	Excitement
Ex Machina	Red	Anger	Red	Fear
Inglorious Basterds	Red	Anger	Red, black	Sadness
Rebel Without a Cause		Anger	Red	Anger
The Shining	Red	Anger	Red	Fear
Deadpool	Red	Anger	Gray, blue	Excitement

Of the 27 films initially labeled, 18 were tagged the same way by the respondents, which represents 66.6% correctness with the original labels. However, there are cases where respondents voted for two opposite emotions in parallel (for example, fear and excitement in "Pirates of the Caribbean").

Nine film excerpts were classified differently by respondents than according to the original labels; for example, "Gladiator," according to the users, evokes the emotion of fear or sadness, even though the original label is excitement. The highest ambiguity in the tags can be seen in Table 14, which shows the results of excerpts where the original label is anger. Only one of the five films with this label was correctly classified as a scene that arouses the emotion of aggression in the viewer.

The concordance of the emotions selected by the respondents compared to the original annotations is higher than the
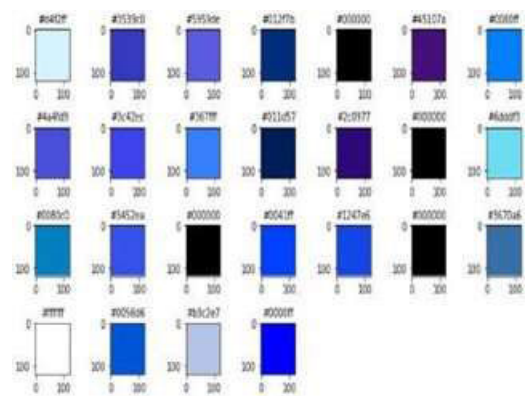

FIGURE 9. Accuracy and model loss for RNN on frames of video fragments after clustering operation into RYB model-based colors and binary division of emotion.

accordance of the colors chosen. For example, in Table 10, "Maleficent" and "Into the Wild" are the only ones of the six where respondents chose green as the color influencing the emotion. However, even in this example, this color was chosen equally to black in the case of the "Maleficent" movie. The highest agreement of annotations of initial colors and those collected through the survey was found with purple and red.

Figures 9 and 10 show sample frames from the film to collect respondents' answers regarding the color present in the scene that influenced their chosen emotion.

The film database shows a direct color-emotion relationship that does not always agree with the respondents' responses. For example, in the original annotations, scenes in which the presence of red predominated have emotion labels of "aggression." In reality, only one in five films with this emotion was classified this way by the respondents On the contrary, films originally classified for the emotions of fear and happiness were labeled the same way by the respondents. These results have a logical explanation − scenes intended to induce fear or happiness usually straightforwardly show this, while emotions such as sadness can take many different forms. Assigning a color to a particular emotion may also not always be correct − for example, in the film "The Shining," the use of large amounts of red was to evoke fear. Even if the
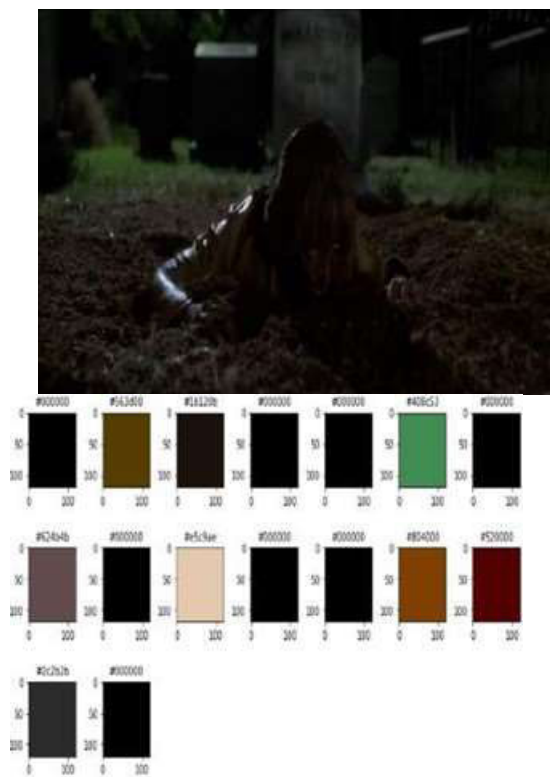
FIGURE 10. Accuracy and model loss for RNN on frames of video fragments after clustering operation into RYB model-based colors and binary division of emotion.

very source of the color red resulted from aggression in the film, it should instill fear in the viewer.

VII. CONCLUSION

Based on the analysis of the results obtained on WikiArt, it can be concluded that reducing the amount of information in the image solely to the color and its location does not significantly change the classification results. Furthermore, none of the three clustering variants improved the algorithm's performance. Moreover, when observing the training of even the best algorithms, instability in training is to be noticed, which may be caused by the limited amount of data. When referring to the literature, the average accuracy of image and emotion classification can vary depending on several factors, such as the type of images used for the classification task, the emotions being predicted, the method used for annotation, the algorithms used for classification, and the evaluation measure employed. Thus, a direct comparison is not possible. However, in general, the accuracy of image and emotion classification can range from around 50% to 70%, or higher, depending on the complexity of the task and the quality of the data. The accuracy of these models can be improved by using more diverse and larger datasets. Overall, the results achieved in the study are in accordance with the state-of-the-art.

Training on film excerpts data, as presented in the "Results" section, differs from experiments on image databases. A division into two categories of emotions was

required here, and other simplified approaches were also used due to the limited amount of data. In contrast to training on image datasets, reducing the amount of information in the video to colors and their location through clustering improved the performance of the algorithms regardless of the type of color clustering scheme. However, clustering to the RYB model-based colors improved the results twice as much as clustering to the image color palette. Such good results may be due to the close relationship noted in Tables 10–14 between the color present in the video excerpt and the tagged emotion.

Understanding how people react to movie excerpts required the creation of an online survey to collect the emotions aroused by scenes in a movie. The results of this survey provided a perspective on looking at emotions and colors. However, the problem of the subjectivity of viewers' emotions is visible. Indeed, it can be solved by increasing the number of people surveyed. Also, a more comprehensive range of viewers' labeling of emotions in film excerpts may help with the contradiction that may exist due to the small sample of survey participants. Another solution is to use movies from less popular productions — a lack of knowledge of the context of the film excerpt may result in the viewer's labeled emotions being the result of only the scene watched. In such a case, observation of the dominant color may also be more focused, resulting in a higher correlation between the color and the emotion assigned.

The above summary may be checked against the literature findings. Modern neuropsychology defines 'color cognition' as a field of science encompassing perception and memory, so there is a clear link between mental representations and the knowledge/memory of objects [49]. Moreover, the basis of color perception is due to Isaac Newton, who, in the 17th century, explained that light is colorless and "the waves themselves are not colored." According to that statement, color is the interpretation of physical phenomena by the human brain based on complex processes, i.e., color has to be perceived inside our brains. Indeed, color perception is of high interest to psychologists, especially in terms of emotion- and behavior-evoking [50]. In a 2020 study, several thousand people responded to a survey that associated color with emotions. They assigned black to sadness, white to relief, red to love, green to contentment, brown to disgust, etc. It should, however, be pointed out that even though 68% of respondents associate red with love, only 36% link brown to disgust. Moreover, some relevant correlations were found by González-Martín et al. at the level of color and emotion, such as the relationships between black and fear. Some colors, i.e., red, violet, and blue, correlate highly with excitement compared to other color-emotion combinations [51]. This shows that one should not expect a uniform association between color and emotion, notwithstanding some colors may have a universal identity. Hence, we expect an adequate recognition of color-emotion association by machine learning when this process is not universal for people, depending on culture, background, age, personality, color preference, etc.

Also, a very interesting notion was brought up in a paper by González-Martín et al. indicating that color is not static when evoking the observer's emotion [51], so this is another factor in emotion evaluation, especially in the case of video.

To summarize, there are several constraints related to our study; some of them are due to the space limit. Among them, insufficient data concerning video annotated excerpts seems the most important one. An answer to this may be to assign emotions to movie excerpts through crowdsourcing. This way, it would be possible to better understand how a movie fragment elicits emotions in people. Moreover, such a manner of data collection would be more reliable in the context of deep model training without concerns that the data would be imbalanced, or the model would be poorly overfitted. Another factor that might influence the outcome of our study is related to emotion convention description, as not all sentiment expressions contained in a prescribed dictionary have the same meaning for all people. So, another task for crowdsourcing could be finding the most frequent sentiment expressions. At the same time, the relationship between color and sentiment may differ between the annotators. However, this may be solved in a similar way by applying an Internet-based survey. Last but not least, there is a plethora of approaches to color analysis in movies. This is one of the future directions that will be pursued, especially as image color analysis translated to the color contained in a video may be problematic. This could be considered a long-term aim. We also believe that the outcome of the crowdsourcing sentiment assignment may identify additional challenges that are related not only to our work, but to more general research.

Finally, when referring to the applicability of such works, studies of the relationship between color and emotion in video excerpts can be used in many areas. The emotions recognized can support both automatic prompts in advertising and marketing, as well as recommendation systems in the context of matching the user's expected emotions. However, this needs large, annotated datasets, even though social crowdsourcing annotations often result in contradictions.

REFERENCES

[1] Y. Liu, Q. Fu, and X. Fu, "The interaction between cognition and emotion," *Chin. Sci. Bull.*, vol. 54, no. 22, pp. 4102–4116, Nov. 2009, doi: 10.1007/s11434-009-0632-2.

[2] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Sci. Inf. Sur Les Sci. Sociales*, vol. 44, no. 4, pp. 695–729, Dec. 2005, doi: 10.1177/0539018405058216.

[3] R. W. Levenson, J. Soto, and N. Pole, "Emotion, biology, and culture," in *Handbook of Cultural Psychology*. New York, NY, USA: The Guilford Press, 2007, pp. 780–796.

[4] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[5] D. Novak, A. Nagle, and R. Riener, "Linking recognition accuracy and user experience in an affective feedback loop," *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 168–172, Apr. 2014, doi: 10.1109/TAFFC.2014.2326870.

[6] E. Schwitzgebel, "Why did we think we dreamed in black and white?" *Stud. Hist. Philosophy Sci. A*, vol. 33, no. 4, pp. 649–660, Dec. 2002, doi: 10.1016/S0039-3681(02)00033-X.

[7] C. F. Hibadullah, A. W.-C. Liew, and J. Jo, "Colour-emotion association study on abstract art painting," in *Proc. Int. Conf. Mach. Learn. Cybern. (ICMLC)*, no. 2, Jul. 2015, pp. 488–493.

[8] B. Ranjgar, M. K. Azar, A. Sadeghi-Niaraki, and S. Choi, "A novel method for emotion extraction from paintings based on Luscher's psychological color test: Case study Iranian–Islamic paintings," *IEEE Access*, vol. 7, pp. 120857–120871, 2019, doi: 10.1109/ACCESS.2019.2936896.

[9] S. Liu and M. Pei, "Texture-aware emotional color transfer between images," *IEEE Access*, vol. 6, pp. 31375–31386, 2018, doi: 10.1109/ACCESS.2018.2844540.

[10] S. Carvalho, J. Leite, S. Galdo-Álvarez, and Ó. F. Gonçalves, "The emotional movie database (EMDB): A self-report and psychophysiological study," *Appl. Psychophysiol. Biofeedback*, vol. 37, no. 4, pp. 279–294, Dec. 2012, doi: 10.1007/s10484-012-9201-6.

[11] T. Lee, N. Lee, S. Seo, and D. Kang, "A study on the prediction of emotion from image by time-flow depend on color analysis," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Las Vegas, NV, USA, Dec. 2020, pp. 747–749, doi: 10.1109/CSCI51800.2020.00141.

[12] S. A. Mohseni, H. R. Wu, J. A. Thom, and A. Bab-Hadiashar, "Recognizing induced emotions with only one feature: A novel color histogram-based system," *IEEE Access*, vol. 8, pp. 37173–37190, 2020, doi: 10.1109/ACCESS.2020.2975174.

[13] Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Affective video content analysis: A multidisciplinary insight," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 396–409, Oct./Dec. 2018, doi: 10.1109/TAFFC.2017.2661284.

[14] J. Wei, X. Yang, and Y. Dong, "User-generated video emotion recognition based on key frames," *Multimedia Tools Appl.*, vol. 80, no. 9, pp. 14343–14361, Apr. 2021, doi: 10.1007/s11042-020-10203-1.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[16] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks," 2014, *arXiv:1410.8586*.

[17] B. Dudzik, J. Broekens, M. Neerincx, and H. Hung, "A blast from the past: Personalizing predictions of video-included emotions using personal memories as context," 2020, *arXiv:2008.12096*.

[18] P. Chua, D. Makris, D. Herremans, G. Roig, and K. Agres, "Predicting emotion from music videos: Exploring the relative contribution of visual and auditory information to affective responses," 2022, *arXiv:2202.10453*.

[19] H. T. P. Thao, B. T. Balamurali, G. Roig, and D. Herremans, "AttendAffectNet–emotion prediction of movie viewers using multimodal fusion with self-attention," *Sensors*, vol. 21, no. 24, p. 8356, Dec. 2021, doi: 10.3390/s21248356.

[20] M. Muszynski, L. Tian, C. Lai, J. D. Moore, T. Kostoulas, P. Lombardo, T. Pun, and G. Chanel, "Recognizing induced emotions of movie audiences from multimodal information," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 36–52, Jan. 2021, doi: 10.1109/TAFFC.2019.2902091.

[21] S. Wang, L. Hao, and Q. Ji, "Knowledge-augmented multimodal deep regression Bayesian networks for emotion video tagging," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 1084–1097, Apr. 2020, doi: 10.1109/TMM.2019.2934824.

[22] D. Brainard, "Color vision theory," in *International Encyclopedia of the Social & Behavioral Sciences*. Oxford, U.K.: Pergamon, pp. 2256–2263, doi: 10.1016/B0-08-043076-7/00666-5.

[23] L. S. Carvalho, D. M. A. Pessoa, J. K. Mountford, W. I. L. Davies, and D. M. Hunt, "The genetic and evolutionary drives behind primate color vision," *Sec. Behav. Evol. Ecol.*, vol. 5, p. 34, Apr. 2017, doi: 10.3389/fevo.2017.00034.

[24] J. Fischbarg, *The Biology of the Eye*. Amsterdam, The Netherlands: Elsevier, 2005. [Online]. Available: https://books.google.pl/books?id=SuKRyC1Q8W4C

[25] T. P. Sakmar and H. T. Rhodopsi, *Encyclopedia of Neuroscience*. Oxford, U.K.: Academic, 2009, pp. 365–372, doi: 10.1016/B978-008045046-9.00922-0.

[26] G. Jordan, S. S. Deeb, J. M. Bosten, and J. D. Mollon, "The dimensionality of color vision in carriers of anomalous trichromacy," *J. Vis.*, vol. 10, no. 8, p. 12, Jul. 2010.

[27] G. A. Agoston, *Color Theory and Its Application in Art and Design*, vol. 19. Berlin, Germany: Springer, 2013.

[28] C. Witzel, J. Maule, and A. Franklin, "Red, yellow, green, and blue are not particularly colorful," *J. Vis.*, vol. 19, no. 14, p. 27, Dec. 2019, doi: 10.1167/19.14.27.

[29] L. Wilms and D. Oberfeld, "Color and emotion: Effects of hue, saturation, and brightness," *Psychol. Res.*, vol. 82, no. 5, pp. 896–914, Sep. 2018, doi: 10.1007/s00426-017-0880-8.

[30] A. Takei and S. Imaizumi, "Effects of color–emotion association on facial expression judgments," *Heliyon*, vol. 8, no. 1, Jan. 2022, Art. no. e08804, doi: 10.1016/j.heliyon.2022.e08804.

[31] S. Bleicher, *Contemporary Color: Theory and Use*. Boston, MA, USA: Cengage Learning, 2012.

[32] F. Leichsenring, "The influence of color on emotions in the Holtzman inkblot technique," *Eur. J. Psychol. Assessment*, vol. 20, no. 2, pp. 116–123, Jan. 2004.

[33] Y. Park and D. A. Guerin, "Meaning and preference of interior color palettes among four cultures," *J. Interior Des.*, vol. 28, no. 1, pp. 27–39, May 2002.

[34] M. Lüscher, *The Lüscher Color Test*. New York, NY, USA: Simon and Schuster, 1990.

[35] (2015). *WikiArt Emotions Dataset*. [Online]. Available: http://saifmohammad.com/WebPages/wikiartemotions.html

[36] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behav. Res. Methods*, vol. 37, no. 4, pp. 626–630, Nov. 2005.

[37] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 83–92.

[38] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cogn. Emotion*, vol. 9, no. 1, pp. 87–108, Jan. 1995, doi: 10.1080/02699939508408966.

[39] A. C. Samson, S. D. Kreibig, B. Soderstrom, A. A. Wade, and J. J. Gross, "Eliciting positive, negative and mixed emotional states: A film library for affective scientists," *Cogn. Emotion*, vol. 30, no. 5, pp. 827–856, Jul. 2016, doi: 10.1080/02699931.2015.1031089.

[40] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cogn. Emotion*, vol. 24, no. 7, pp. 1153–1172, Nov. 2010, doi: 10.1080/02699930903274322.

[41] B. Zupan and M. Eskritt, "Eliciting emotion ratings for a set of film clips: A preliminary archive for research in emotion," *J. Social Psychol.*, vol. 160, no. 6, pp. 768–789, Nov. 2020, doi: 10.1080/00224545.2020.1758016.

[42] (2017). *ColorThief*. [Online]. Available: https://pypi.org/project/colorthief/

[43] K. P. Sinaga and M.-S. Yang, "Unsupervised k-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020.

[44] T. Ciborowski, S. Reginis, D. Weber, A. Kurowski, and B. Kostek, "Classifying emotions in film music—A deep learning approach," *Electronics*, vol. 10, no. 23, p. 2955, Nov. 2021, doi: 10.3390/electronics10232955.

[45] *OpenCV Python*. Accessed: May 2023. [Online]. Available: https://pypi.org/project/opencv-python

[46] *Machine Learning in Python*. Accessed: May 2023. [Online]. Available: https://scikit-learn.org/stable/

[47] Y. A. Ng, "Feature selection, L_1 vs. L_2 regularization, and rotational invariance," in *Proc. 24th Int. Conf. Mach. Learn. (ICML)*, Banff, AB, Canada. Stanford, CA, USA: Stanford Univ., Computer Science Department, Jul. 2004, doi: 10.1145/1015330.1015435.

[48] B. Juba and H. S. Le, "Precision-recall versus accuracy and the role of large data sets," in *Proc. 33rd AAAI Conf. Artif. Intell.*, vol. 30, 2019, pp. 4039–4048, doi: 10.1609/aaai.v33i01.33014039.

[49] J. Davidoff, *Cognition Through Color*. Cambridge, MA, USA: MIT Press, 1991.

[50] K. Cherry, "Color psychology: Does it affect how you feel? How colors impact moods, feelings, and behaviors," *Cogn. Psychol.*, pp. 1–8. Accessed: May 2023. [Online]. Available: https://www.verywellmind.com/color-psychology-2795824

[51] C. González-Martín, M. Carrasco, and G. Oviedo, "Analysis of the use of color and its emotional relationship in visual creations based on experiences during the context of the COVID-19 pandemic," *Sustainability*, vol. 14, no. 20, p. 12989, Oct. 2022, doi: 10.3390/su142012989.

ALEKSANDRA WĘDOŁOWSKA received the B.Sc.Eng. degree, in 2021, and the M.Sc. degree in biomedical engineering with a specialization in artificial intelligence, in 2022. She was a Student of the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology. She created an application that aimed to help Alzheimer's and dementia patients during the B.Sc.Eng. degree. The thesis was about developing a machine-learning algorithm for classifying emotions based on colors used in movie fragments. Her main research interests include computer vision, creating mobile applications, and connecting artificial intelligence technologies with everyday activities.

DAWID WEBER was born in Toruń, in 1992. He received the degree from the Department of Multimedia Systems, Faculty of Electronics, Telecommunications and Informatics, in 2017. The subject of the engineering thesis concerned the guitar sound effects processor applied to a mobile device, while the M.Sc. thesis was devoted to music band recording using various microphone techniques, along with a subjective assessment of the resulting sound. His research interests include using machine learning techniques and algorithms to predict emotions from music and films.

BOŻENA KOSTEK (Senior Member, IEEE) is currently a Professor with the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology (GUT), Poland. She is also a Corresponding Member of the Polish Academy of Sciences. She has supervised more than 300 master's and engineering theses and 21 Ph.D. theses. She has also led a number of research projects. She has presented more than 600 scientific papers for journals and at international conferences. She has also published three books related to multimedia applications. Her main research interests include acoustics, psychoacoustics, multimedia, music information retrieval, cognitive and behavioral processing, as well as applications of machine learning to the mentioned domains. She is a fellow of the Audio Engineering Society and the Acoustical Society of America. She was a recipient of many prestigious awards for research, including those of the Prime Minister of Poland (twice), the Ministry of Science, and the Polish Academy of Sciences. She was the Editor-in-Chief of the *Journal of the Audio Engineering Society* as well as an Associate Editor of IEEE/ACM Transactions on Audio, Speech, and Language Processing and the Guest Editor of *JASA*.

• • •