IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 12, 2020, accepted August 20, 2020, date of publication September 1, 2020, date of current version September 16, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3020918*

# Categorizing Color Appearances of Image Scenes Based on Human Color Perception for Image Retrieval

**ANIZA OTHMAN[1], TENGKU SITI MERIAM TENGKU WOOK[2], AND FAIZAN QAMAR[2]**

[1]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Durian Tunggal 76100, Malaysia
[2]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Selangor 43600, Malaysia

Corresponding author: Aniza Othman (aniza.othman@googlemail.com)

This work was supported in part by the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, for providing equipment, facilities and financial support Under Research Development Scheme (PRGS/2/2019/ICT04/UKM/02/1) (Universiti Kebangsaan Malaysia).

**ABSTRACT** Institutions that possess certain collections of digital image libraries, such as museums, are progressively interested in making such collections accessible anytime and anywhere for any Image Retrieval (IR) activities namely browsing and searching. Many researchers have shown that IR methods, in filtering images based on their features such as colors, would provide better indexing and can be able to deliver/provide more accurate results. The color composition of an image, e.g. color histogram has proven to be a powerful feature that can be analyzed and used for image indexing because of its robust standardization of image transformation such as scaling and orientation. In this research, the efforts in narrowing the gap between low relevancy human text descriptions for Malaysian users and image scene color appearances have been brought into attention. The methods are first, to investigate the color concepts and color appearance descriptions of a scene and secondly, to identify a set of ground-truth images for each color appearance category. Psychophysical experiments are conducted to determine a collection of ground-truth images that effectively match five color appearance descriptions for image scenes in accordance with human judgement and perception. The results of the experiments are presented together with the inter-rater agreement analysis. These descriptions that are commonly queried by humans are the following keywords, *Bright, Pastel, Dull, Pale,* and *Dark*. The agreement analysis indicates that the *Bright* category is the most comprehensible by humans and subsequently followed by the *Pastel* and *Dark* categories. *Dull* and *Pale* categories, on the other hand are fairly understood by humans. All the images involved in this research are landscape painting collections from the internet and they are used for academic purposes only. The results show the top ten ground-truth images for each category that encapsulates a high level of agreeability between humans.

**INDEX TERMS** Color appearance description, color concept, color features, human judgement, human agreement analysis, psychophysical experiments.

## I. INTRODUCTION

In today's modern world, most images are in a digitalized form and kept in digital libraries for better management. With the growth of information and communication technology, digital image collection holders have been increasingly interested in making their collections available for any IR activities such as browsing and searching. Recently, designing a search image mechanism based on user requirements has become an important and critical challenge [1]–[3]. Depending on the application system, image searching can be done using either a text description query or image query and the retrieval process of the images is done by using their index. However, due to humans being subjective and boasting different creative natures, the user's satisfaction is in doubt, because this image indexing method uses a manual text description approach that can be rather rigid. For that reason, an indexing method using image content such as colors is likely to be preferred as it can be done automatically to each image in digital libraries and provide a better set of results [4].

Colors are always said to be the easiest and most reliable feature captured by the human eye. Color distribution in an image can be represented globally as a descriptor in the form of a color histogram [5], [6]. It has been observed

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

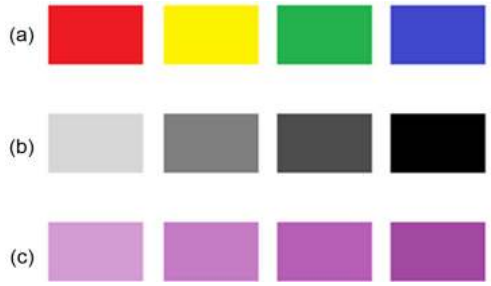



**FIGURE 1.** (a) Example of hues: red, yellow, green, blue (b) Example levels of lightness (light-dark) (c) Example levels of saturation for the same hue (low – high).

that the color histogram approach is the most preferable method in many IR applications [7]–[9]. Color histogram can be generated and calculated through the quantization process of the color model. According to Saima *et al.* [10] in their work, an appropriate decision on interval size in quantization is needed in reducing the quantization affection error. Similarly, color quantization is a process to reduce the number of colors in an image without degrading its quality. Thus, with a suitable color model for color representation and quantization level, the generated color histograms can be used effectively for image indexing [11]–[13]. In our previous work, A. Othman *et al.* [12] studied on the suitable quantization level for generating an optimum color histogram for all five categories of color appearance in this study. We used the CIELab color model as this model represents colors similar to how humans perceive colors using its color attributes, *hue, saturation,* and *lightness*.

Color features are robust to noise, image degradation, size changes, resolution and orientation [14]–[17]. However, colors are often a difficult matter to properly describe, not only in technical terms such as *lightness, saturation,* and *hue*, but also to express in simple words such as *bright, dark, vivid, pale, dull, pastel*. These descriptive words are hard to use accurately at first, let alone unfamiliar color descriptive words such as *pure, muted, ablaze, fiery*, *restrained, brash, jazzy* etc. The MacMillan English Dictionary has listed more than eighty general words that are used to describe colors [18]. The meaning of these words is regularly but vaguely used, as the words are misunderstood in usage.

Visual psychology research shows that when humans observe an image, not all of the contents are of the same interest [19]. Colors decoded in the human mind might be different from others. It is a subjective sensation that includes physics, biology, and psychology etc. In perceptual color theory, color consists of three attributes known as *hue*, *brightness* or *lightness* and *saturation* [20]. *Hue* is defined as the names of color such as red, yellow, green, and blue. Different hues are caused by different wavelengths of light. *Brightness* or *lightness* is meant by the grayscale ranging from black (lowest) to white (highest) and saturation means the amount of pure color that exists. Different variations of these attributes are shown in Fig. 1.

Perceptual colors depend on certain criteria such as the experience of the viewer, and the conditions of the color

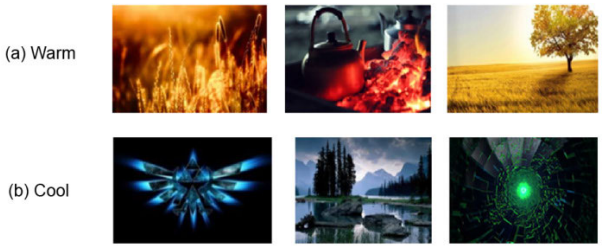



**FIGURE 2.** An example of images based on art concept – (a) *Warm* images (b) *Cool* images. Images are retrieved from Google Image Search by keywords 'warm images' and 'cool images'.


**FIGURE 3.** An example of four images that shows the different appearance of colors. Images are from Google Image search.

or other colors around it. Humans describe colors based on both the color component (true colors such as red, blue, green etc.) and the intensity component (amount of lightness and saturation). In a color IR application, image retrieved by users is accomplished according to their specifications on what they want or acquire, that could be based upon so many concepts. For instance, *Warm* and *Cool* are art concepts. According to psychologist researchers, *Warm* images contain warm colors such as yellow or red in their content, whereas *Cool* images are images with cool colors such as blue or green [21]. Some example images with these concepts are illustrated in Fig. 2. They were retrieved using Google search engine. The effective IR is based on the concept that needs efficient and meaningful indexing to make sure transfer of image information from the database to a user is as accurate as possible [22].

Based on the work by Gabbouj *et al.* [23], the intensity of colors among image appearance concepts such as bright, pastel, or dark can be quantified through human visual perception. Because of human subjectivity, what is described could therefore vary from one human to another. So far, however, there has been minimal discussion about human descriptions for the color appearance of image scenes. Fig. 3 shows four images with different appearance of colors. *What appearance descriptions can be given to describe the entire image individually?* This task can help to envision how to group all the hues or colors in the images, which consists of various levels of saturation and lightness in only one word.

In addition to the misunderstood meaning of color descriptions, humans, especially Malaysian users, tend to think that images comprising of different colors are called colorful images [24].

In color theory, colorful or colorfulness means high saturation or high chroma or bright colored images. For Malaysian users, the term *multicolored* is unfamiliar, but it is the most suitable word to describe images with many colors.


**FIGURE 4.** Example of images randomly from Google Image Search – (a) *Colorful* images using the keyword 'colorful images' (b) *Multicolored* images using the keyword 'multicolored images'.

A *multicolored* image could also appear as a pastel or a bright image, etc. However, the *multicolored* appearance is not discussed in this paper. Fig. 4 (a) and 4 (b) shows some colorful and multicolored images, respectively, browsed using Google Image search.

In this section, we give a general overview of the study of the concept of image scenes based on human color perception. The remaining part of this paper is organized as follows. Section 2 explains the secondary researches related to the study. Section 3 highlights the contribution of this study. Section 4 discusses the methods used to obtain the ground-truth images for five color appearance categories. Section 5 shows the experiment analysis and discussion. Section 6 discusses the results and conclusion of the paper. Finally, section 7 highlights the limitation and the future works.

## II. BACKGROUND

As mentioned above, in a color IR application, image retrieved by users is accomplished according to their specifications on what they want or acquire, which could be based upon so many concepts. Each concept must have just the right descriptors to make retrieved images relevant. For instance, IR systems that retrieve color images based on the whole appearance of colors should have color appearance descriptors [19], [25]–[27].

In general, the descriptors should at least reduce the gap between high-level and low-level concepts. High-level concepts refer to descriptions that humans use when they want to find certain images, e.g. ''Find me a painting that has a *pastel* vase in it''. On the other hand, low-level concepts are visual features of images, e.g. color that needs to be extracted and represented in bits or bytes to allow computers to understand, analyze and process into some kind of formulation that can help IR queries.

Current IR systems or machines on color intensity use mathematical algorithm without considering collective human opinions. However, in an IR system, to describe or to develop color appearance descriptors of an image scene using low-level features is a difficult task to achieve. In order to build a system that performs similar to how humans look at colors, it is prudent to have empirical investigations of human color perception.

In the present study, the revealed perceptual and semantic relation between descriptions and color appearances would provide the basis for creating such simulation systems of human color perception and semantic processes. For such a system, it is necessary to have experience and insight into human color perception and its relationship with their description [28], [29]. Although everyone has different color recognitions, it can be generalized by group investigations.

Several psychophysical experiments related to human color perception have been carried out to investigate the human description and human judgment of colors. Some psychophysical experiments have been conducted on how humans rate the similarity of images based on colors and texture features [30], [31]. The study in [32] proposed a method for evaluating and classifying a color image concerning a pair of emotions such as *cool-warm*, *static-dynamic* and *heavy-light* categories. In the study, before classification, the representative color images were chosen for each emotion by human subjects and stored as cases.

Van de Weijer *et al.* [33] investigated the use of color names in images from real-world applications such as images taken under varying illuminants, colored shadows or interreflections. In their work, they used humans to manually label four categories of object images from the auction website *eBay* to collect human varying descriptions on colors, such as, *dark red* and *pale yellowish pink* which refers to a single color name.

On the other hand, some researches focused on the retrieval of colored images using dominant colors to detect objects in images [34] or any similarities of image color content [35] or by using the relationship between colors such as the contrast of *hues*, *light-dark* contrast, *cool-warm* colors etc. Lay and Guan proposed an art concept-based retrieval engine in [36]. It is stated that the images can be retrieved based on art concepts such as *perceptual-harmony*, *perceptual-warmth*, *analogous-harmony*, *saturated/dull* color. However, this work has not included any human judgement. Besides, art concept-based IR has been applied in a search engine such as Google.

Fig. 5 shows some images retrieved from Google using its *Search by Image* tool for two sample images. Fig. 5(a) shows a query image that was interpreted in Google as *warm* paintings; hence all retrieved images were quite similar in colors. Fig. 5(b) shows a query image that has no interpreted information on the internet. All retrieved images also seem to display similar colors. However, query by image is not the

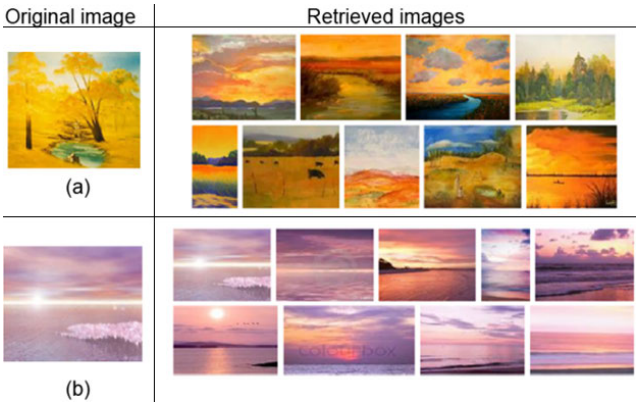



**FIGURE 5.** Two query images (a) and (b) and their retrieved images from Google Search by Image showing visually similar images by color.

scope of this research. We simply want to share the color similarity of query-retrieved images in Google Image Search Engine. Both image retrievals were based on hue distributions. So far, as mentioned before, there has been very little discussion about human descriptions for the intensity-based color appearance of the image scene regardless its hue distributions.

Tai *et al.* [37] have carried out some studies in generating 26 adjectives of color descriptions such as *Fresh, Luxury*, *Romantic, Passion*, *Gentle* etc. Focused groups research and scholars were involved in the experiments and the analysis was done on web colors. Giragama *et al.* [38] studied on how Japanese and Sinhala native viewers describe differences in color tone (modifiers) such as *bright, vivid, strong, dark, pale,* and *dark* in their native languages. In this study, the researchers compare a color with various levels of lightness and saturation and had asked the viewers to describe the color individually. In their preliminary elicitation, they came out with English descriptions that were most frequently used and constituted to a high proportion of responses, which were *Bright, Vivid, Strong, Dull, Pale*, and *Dark.*

## III. CONTRIBUTION

In this study conducted on Malaysian users, we used the English word descriptions from [38] which are *Bright, Dull, Pale, Dark* and based on our prior knowledge, and we include *Pastel* for another description. We have conducted an experiment on eight Malaysian Color Experts for triangulation of the results. The Triangulation technique can be used to cross-check the results of the same experiment but with different methods [39]. In the experiment, they were asked to select and describe the color appearance of image scenes. Based on our finding, the dataset images obtained were almost the same. Additionally, *Bright, Dull, Pale, Dark, Pastel* and *Colorful* have been identified to be the most commonly used word descriptions for Malaysian users. For searching/browsing activities, these descriptions should also be used as search keywords. More information about the experiment can be found in A. Othman *et al.* [24].

Ground-truth images are important in our research to ensure the results of the study is valid for current and

**TABLE 1.** Definition of word description for color appearance given by Macmillan dictionary and dictionary.com.

| Word Description | Definition by Macmillan Dictionary and Dictionary.com |
|---|---|
| Bright | strong, brilliant and vivid colors |
| Pastel | soft and subdued colors |
| Dull | having very little depth of color; lacking in richness or intensity of color, not bright, intense, or clear; dim |
| Pale | lightly-colored or lacking in color a low degree of chroma, saturation, or purity; approaching white or gray not bright or brilliant; dim |
| Dark | not pale, approaching black in hue |

**TABLE 2.** Respondent biography.

| Profession | Gender | Field |
|---|---|---|
| Student (53) | Female (54) | Designing (10) |
| Worker (24) | Male (27) | Engineering (11) |
| Pensioner (4) | | ICT (42) |
| | | Management (5) |
| | | Business (5) |
| | | Education (5) |
| | | Medical (3) |

future research. As far as the author's work indicates, there is not yet a dataset image that has been verified (ground-truth) as knowledge for set data *Bright, Dull, Pale, Dark* and *Pastel.* Table 1 shows the definition of descriptions for each color appearance of landscape paintings used as our dataset. These landscape paintings were browsed using Google Image search engine with keyword 'landscape paintings.' Some of the images were non-commercial reuse and some were downloaded from various websites regardless of what the websites were intended for and they were used strictly for academic purposes only.

Landscape paintings are a depiction in art that feature views of nature, including seascapes, cityscapes, and waterscapes. It captures sceneries from various locations of life and involves a complicated choice of colors and shading to mimic changing seasons or time of the day. Thus, it is suitable for the research, since the research emphasizes more on the harmonious blend of colors perceived as a whole. The contribution of this study is the dataset of ground-truth landscape painting images for five categories of color appearance.

## IV. RESEARCH METHODS

This study has conducted a psychophysical experiment in determining the ground-truth images. The experiment was a participatory survey on human judgment, where respondents categorized 200 landscape paintings. The subjects or the respondents were a collection of 81 Malaysian users of various backgrounds, as shown in Table 2. We believe that these 81 respondents involved, should be enough to produce an initial positive result for a participatory experiment related to human judgement.

Respondents were students, workers and pensioners from different fields such as art and design, engineering, Information Communication and Technology (ICT), administration, business, educators and the medical field.

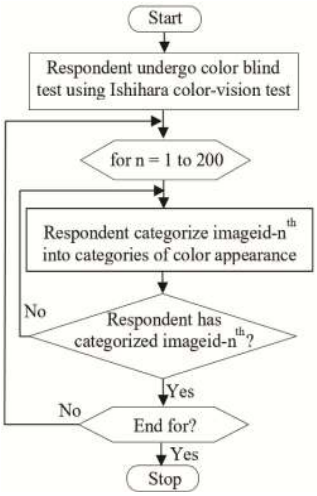



**FIGURE 6.** Flowchart shows the process in the psychophysical experiment.

These respondents were selected for their availability and convenience. They were aged between 20 – 60 years old; 54 of them (66.7%) were female and 27 (33.3%) were male.

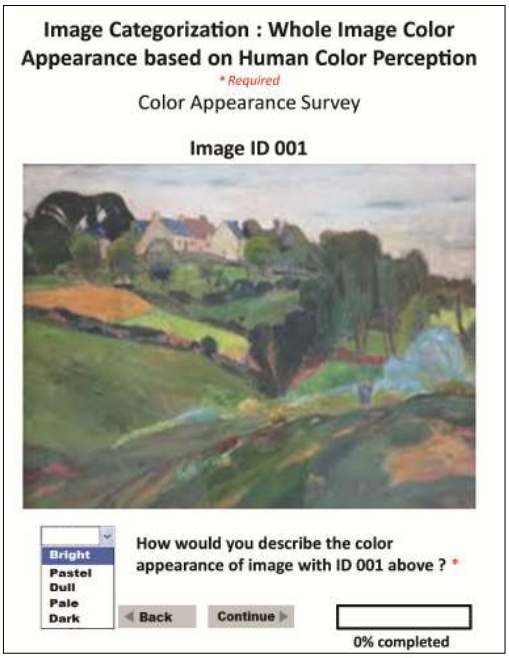
The reason behind the variety in the background of respondents was, in our opinion, the many internet users of Malaysia that are actively searching for images. Web browsing could be done by anyone regardless of their age, work, gender, or level of education. Therefore, in this study, we attempt to get one initial collective result first, and subsequently, in the future, a series of experiments will be done again based of a specified criteria for instance *careers,* and learn whether the initial result can possibly be generalized or otherwise.

The experiment was based on the steps as shown in the flowchart in Fig. 6. The respondents were first tested for color-blindness using Ishihara color-vision test [31] and took the experiment one at one time using the same computer, in the same room, under researcher supervision. This was to ensure that the respondents took the test in the most similar of conditions as possible. The size of the images was varied, but they were shown encompassing the entire computer screen, respectively. All images were labelled in numerical order as 001 - 200.

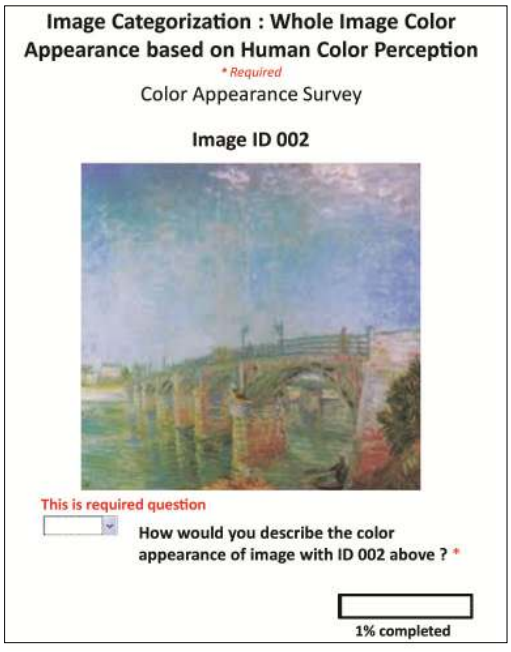
Respondents were required to choose how they perceive the color appearance of the image scenes. The categories were *Bright, Pastel, Pale, Dull* and *Dark*. The respondents must give a response before they can proceed to the next image to ensure all respondents have answered the questionnaires. In general, each respondent took 30 to 40 minutes to finish an experiment. Fig. 7(a) and 7(b) shows examples of the questionnaire for image-id 001 and 002, respectively. The data obtained from the respondents were collected and put directly into one file.

## V. EXPERIMENT ANALYSIS AND DISCUSSION

After the psychophysical experiment was conducted, the feedback data was compiled and analyzed. The important analysis needed for this survey was the frequency distribution of five color appearance categories for every




(a). The questionnaire for the image-id 001




(b). The questionnaire for the image-id 002

**FIGURE 7.** (a). The questionnaire for the image-id 001. (b). The questionnaire for the image-id 002.

200 images and measure of agreement between respondents. Table 3 shows the frequency distribution for 20 images taken from the experiment as an example. These 20 images were the first 10 images of labelled image-id 001 until image-id 010, and the last 10 images of labelled image-id 191 until image-id 200. Based on frequencies, some information can be extracted. Here, the landscape painting image with id 001 was categorized as *Bright* by 15 respondents, *Pastel* by 45 respondents, *Dull, Pale* and *Dark* by 12,

**TABLE 3.** Distribution of respondents by image id and color appearance categories for the first and the last 10 images.

| image-id | Bright | Pastel | Dull | Pale | Dark |
|---|---|---|---|---|---|
| 001 | 15 | **45** | 12 | 3 | 6 |
| 002 | 9 | **42** | 21 | 3 | 6 |
| 003 | **60** | 15 | 3 | 0 | 3 |
| 004 | **81** | 0 | 0 | 0 | 0 |
| 005 | 12 | **27** | 21 | 21 | 0 |
| 006 | 3 | 24 | **27** | **27** | 0 |
| 007 | 3 | **36** | **36** | 6 | 0 |
| 008 | **54** | 15 | 0 | 9 | 3 |
| 009 | **60** | 18 | 0 | 0 | 3 |
| 010 | 3 | 18 | **21** | 18 | **21** |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 191 | 0 | 12 | **33** | 24 | 12 |
| 192 | 0 | 9 | 24 | 9 | **39** |
| 193 | 0 | 18 | 21 | **39** | 0 |
| 194 | 3 | 24 | **27** | **27** | 0 |
| 195 | 18 | **42** | 15 | 6 | 0 |
| 196 | **66** | 12 | 3 | 0 | 0 |
| 197 | **60** | 21 | 0 | 0 | 0 |
| 198 | 18 | **27** | 9 | **27** | 0 |
| 199 | **69** | 9 | 0 | 0 | 3 |
| 200 | 9 | 15 | **39** | 18 | 0 |

**TABLE 4.** Percentage of images classified by human in each category.

| Appearance | Number of Images categorized by human |
|---|---|
| Bright | 53 images or 26.5% |
| Pastel | 46 images or 23.0% |
| Dull | 41 images or 20.5% |
| Pale | 10 images or 5.0% |
| Dark | 35 images or 17.5% |
| *Conflict | 15 images or 7.5% |

*Based on human color perception survey, 185 images have been categorized into their categories based on their highest frequency. 15 images have conflict color appearances.

3 and 6 respondents, respectively. Based on the responses, *Pastel* appearance has the highest frequency, therefore, image-id 001 could be said to be perceived as *Pastel* by most people.

After examining the distribution and identifying the category with the highest frequency for all images, we have obtained the results, as shown in Table 4. This table shows the percentages of images in each category based on respondents' color perception. A total of 185 images have been categorized into their categories based on their highest frequency; 53 images or 26.5% were categorized as *Bright*, 46 images or 23% as *Pastel*, 20.5%, 5%, and 17.5% images were categorized as, *Dull, Pale,* and *Dark* respectively. It is also observed that there were 15 (7.5%) images that belong in a 'conflict' category of color appearance, as they have the same highest frequencies in more than 1 category. Thus, these images have been excluded from the analysis in determining ground-truth images. Examples of *conflict* images shown in Table 3 are image-id 006, 007, 010, 194 and 198. Image-id 006 was excluded from having the same highest frequency in two categories: *Dull* and *Pale*. This might possibly have an interesting explanation that needs to be studied on how the respondents categorized the image as they were but for now, that is not in the scope of this paper. The highest frequency obtained by a single image indicates the percentage of respondents' agreement to agree in classifying the image into a category. Some images may have a higher or lower percentage of respondents' agreement than others.

**TABLE 5.** Image percentage based on respondents' agreements.

| Color Appearance | Total Images | Images by % agreement | | |
|---|---|---|---|---|
| | | Over 75% | 50%-75% | 25%-50% |
| Bright | 53 (26.5%) | 22 (11%) | 22 (11%) | 9 (4.5%) |
| Pastel | 46 (23%) | 1 (0.5%) | 22 (11%) | 23(11.5%) |
| Dull | 41 (20.5) | 0 | 9(4.5%) | 32(16%) |
| Pale | 10 (5%) | 0 | 6 (3%) | 4(2%) |
| Dark | 35 (17.5%) | 4 (2%) | 13 (6.5%) | 18(9%) |

**TABLE 6.** Percent agreement (*%A*) for ten images for each category.

| Bright | | Pastel | | Dull | | Pale | | Dark | |
|---|---|---|---|---|---|---|---|---|---|
| id | *%A* | id | *%A* | id | *%A* | id | *%A* | id | *%A* |
| 004 | 100 | 031 | 81.48 | 161 | 62.96 | 060 | 66.67 | 037 | 96.30 |
| 015 | 100 | 064 | 70.37 | 050 | 59.26 | 084 | 59.26 | 056 | 88.89 |
| 017 | 100 | 107 | 70.37 | 153 | 59.26 | 176 | 55.56 | 019 | 85.19 |
| 089 | 100 | 180 | 70.37 | 081 | 55.56 | 086 | 51.85 | 067 | 77.78 |
| 101 | 100 | 093 | 66.67 | 042 | 51.85 | 185 | 51.85 | 110 | 74.07 |
| 140 | 100 | 186 | 66.67 | 063 | 51.85 | 152 | 51.85 | 018 | 70.37 |
| 079 | 96.3 | 025 | 62.96 | 076 | 51.85 | 193 | 48.15 | 021 | 70.37 |
| 090 | 96.3 | 065 | 62.96 | 144 | 51.85 | 070 | 44.44 | 022 | 70.37 |
| 033 | 92.6 | 092 | 62.96 | 190 | 51.85 | 109 | 40.74 | 023 | 70.37 |
| 035 | 92.6 | 172 | 62.96 | 032 | 48.15 | 087 | 37.04 | 165 | 70.37 |

A total of 53 images are categorized as *Bright*, however, percentages of agreement in each image differs. For example, image-id 004 had a very high levels of agreement which is 100% among all 81 respondents saying that image was *Bright*. Compared to image-id 008, this image has a lower percentage of agreement of 67% with 54 of 81 respondents saying that it was *Bright*.

If we could set the agreement percentage of these images into three ranges, for instance; i) over 75% ii) 50%–75% and iii) 25%–50% for each color appearance category, we would have groups of images with their range of percentage agreement as shown in Table 5. Take the *Bright* category for example. A total of 26.5% of images have been categorized as *Bright* in general. 11% of them were agreed by more than 75% agreement as *Bright*. Another 11% of *Bright* images have 50%–75% agreement and 4.5% of *Bright* images were received between 25%–50% of respondents' agreement. The higher the percentage of agreement, the more precise that the image is to be considered as a ground-truth image.

For the determination of the ground-truth images, we have decided to select ten images for each category with the highest percentage of respondents' agreement. To do that, we ranked all images according to their percentage agreement individually for each appearance. The top ten images for each category have been identified.

Table 6 shows the ranking information. When humans are involved as a part of the measurement procedure, it is imperative to ensure that the reliability and $AC_1$ agreement coefficient gives a value between 0 and 1; 0 being the lowest level of agreement and 1 being the highest. Further information on $AC_1$ agreement coefficient can be found in [40].

**TABLE 7.** $AC_1$ values.

| Set of images | $*AC_1$ Values |
|---|---|
| Total of 200 data set images | 0.266 |
| Total of 50 Images - 10 highest-ranking images for each category | 0.453 |
| 10 high ranking images in *Bright* | 0.956 |
| 10 high ranking images in *Pastel* | 0.411 |
| 10 high ranking images in *Dull* | 0.256 |
| 10 high ranking images in *Pale* | 0.264 |
| 10 high ranking images in *Dark* | 0.575 |

**TABLE 8.** Landis and Koch's benchmark scale (1977).

| Agreement Value | Level of Agreement |
|---|---|
| < 0.0 | *Poor* |
| 0.0 to 0.20 | *Slight* |
| 0.21 to 0.40 | *Fair* |
| 0.41 to 0.60 | *Moderate* |
| 0.61 to 0.80 | *Substantial* |
| 0.81 to 1.00 | *Almost Perfect* |

As stated previously, this research study involves 81 numbers of respondents (raters), 200 of images (participants) and five categories of scene color appearances. The $AC_1$ has been calculated for all 200 images, for all 10 high ranking images of all categories (50 images) and for 10 high ranking images in each category (10 images). Table 7 shows the values of Gwet's agreement coefficient, $AC_1$.

The discussion will start with justification for each $AC_1$ values related to the consistency of the observation by the respondents for overall categories as well as for each category. Table 8 describes the benchmark scale used for the discussion. This benchmark scale was proposed by Landis and Koch (1977) and has been recommended as a useful guideline for practitioners [41]. From this table, the extent of agreement can be qualified as *Poor, Slight, Fair, Moderate, Substantial,* and *Almost Perfect*, depending on the magnitude of $AC_1$. An $AC_1$ value between 20% and 40% indicates a *Fair* agreement level, between 40% and 60% indicates a *Moderate* agreement level, while ranges of values 60% − 80%, and 80% − 100% indicate *Substantial* and *Almost Perfect* agreement levels, respectively. As shown in Table 7, the overall $AC_1$ for all 200 images are 0.266. According to the benchmark scale, the consistency of the respondent agreement is *Fair*. According to Wong *et al.* [42], for social science research, a lower level of agreement is acceptable.

However, $AC_1$ value of all 50 data of set images (which have the highest percentage agreement) selected from the top ten of each category is 0.453. This shows a more consistent respondent agreement. Therefore, these 50 data of set images are distinguished as ground-truth images that could be used in the next experiment.

If the 50 data of set images were further specified according to the categorical labels *Bright, Pastel, Dull, Pale,* and *Dark*, their $AC_1$ values would be 0.956, 0.411, 0.256, 0.264, and 0.575, respectively. All the values lie between the range of *Fair* and *Almost Perfect* that could be accepted with justification.

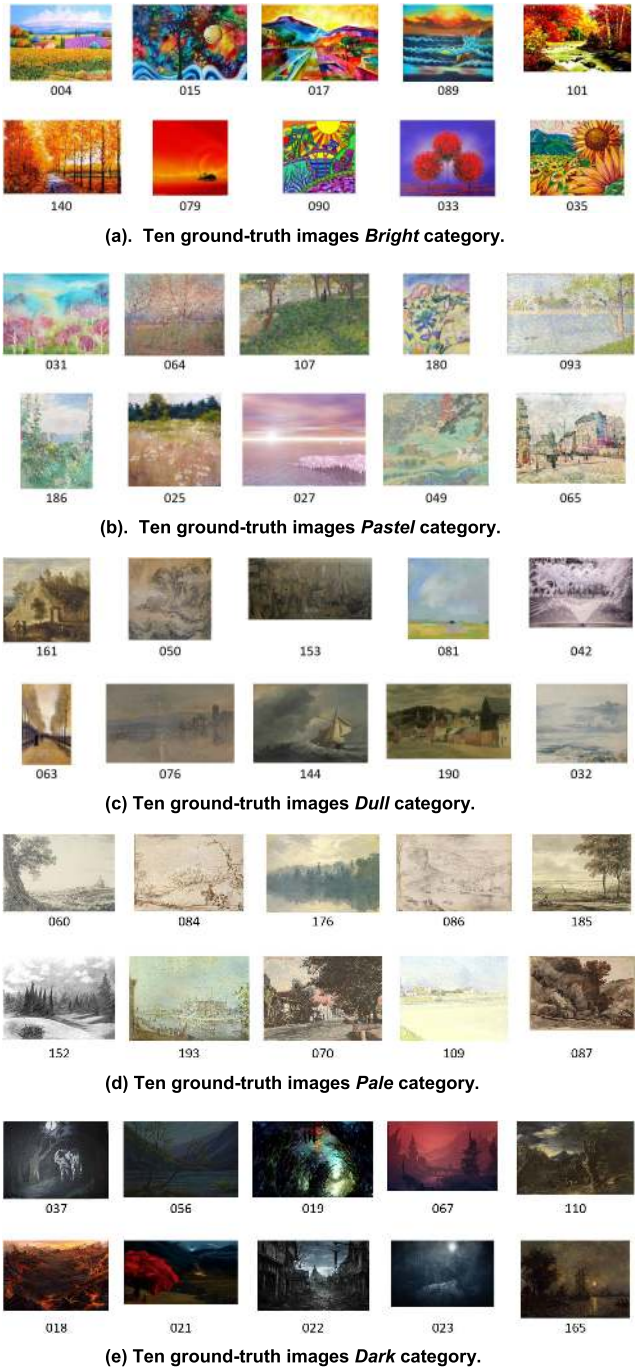



**FIGURE 8.** (a). Ten ground-truth images *Bright* category. (b). Ten ground-truth images *Pastel* category. (c) Ten ground-truth images *Dull* category. (d) Ten ground-truth images *Pale* category. (e) Ten ground-truth images *Dark* category.

Based on Landis and Koch's benchmark scale, the *Bright* categories have an *Almost Perfect* rating for its $AC_1$, which shows that humans are able to better categorize *Bright* images without significant doubt. It could be concluded from these results that most humans agree and have a better understanding of what is perceived as *Bright* images.

This is opposed to the *Dull* and *Pale* category, where respondents have a difficult time differentiating between the two and categorizing images accordingly. This could be

attributed to the lack of clarity for humans in differentiating images that are *Dull* or *Pale*. For images in the *Pastel* and *Dark* category, they have a *Moderate* respondent agreement, and it could be said that human understanding for these two categories is of a *Moderate* level.

This finding is interesting in a sense and should be further investigated; thus subsequent surveys could be conducted to better expound the differing criteria of all categories and how they are understood by humans especially for the obvious ones such as the *Dull* and *Pale* category. This ambiguity will be the essence of future research.

## VI. RESULT AND CONCLUSION

From the statistical analyzed data, 10 images with a high percentage of the human agreement were selected as ground-truth samples for *Bright, Pastel, Dull, Pale* and *Dark* categories. The images are as shown in Fig. 8 (a)–(e), respectively.

Fig. 8(a) shows the top 10 *Bright* images that have an *Almost Perfect* respondents' agreement and have strong, brilliant, and vivid colors. Fig. 8 (b) and 8(e) show top 10 *Pastel* and *Dark* images that have *Moderate* respondents' agreement, respectively. The *Pastel* category shows images with soft and subdued colors, and *Dark* images are not pale but instead approaching black in hue. On the other hand, Fig. 8(c) and Fig. 8(d) shows top 10 images of *Dull* and *Pale* categories that have *Fair* respondents' agreement, respectively. The *Dull* category shows images lacking in richness or intensity and/or dim lightly colored. The *Pale* category shows images that have low chroma, approaching white or gray in appearance.

The results obtained in this study are based on 81 respondents, while the landscape painting images used in the experiment are those that exist in our collections at the moment. However, humans, as we have discussed before, are subjective. It is likely through a series of experiments that we would finally be able to collect a set of images with higher respondents' agreement for all categories and reduce the semantic gap of color appearance meaning from human color perception.

The results can be generalized, and we believe the outcome could be better. Our work in [24] had discussed a triangulation technique that shows similar dataset results obtained using another method. We are positive that the ground-truth images identified through this study should be reliable for our future work, especially in IR applications for landscape paintings.

In conclusion, in this paper, a set of the ground-truth images has been determined based on color appearance concepts. Five categories of color appearance for landscape paintings that are most commonly queried by humans using keywords have been selected for investigation. These categories are the color appearance concepts, which are described as *Bright, Pastel, Dull, Pale,* and *Dark*.

An experiment on human color perception and judgment or psychophysical experiment has been performed for Malaysian users. They have categorized 200 landscape paintings according to how they perceive the color appearance of the entire scene and the experiment feedback data had been analyzed. For each category, the top 10 images in the ranking have been determined and their $AC_1$ agreement coefficient was calculated to observe the consistency of observation by the respondents.

The results have then been verified using Landis and Koch's benchmark scale. The results showed that the *Bright* category is the most understandable by humans subsequently followed by the *Pastel* and *Dark* categories which were moderately understood by a human. The meaning of *Dull* and *Pale* categories, on the other hand were fairly understood.

Most IR applications are based on true colors or hues distributions such as *Warm*, and *Cool*, etc., but have little to no discussions on an intensity-based color appearance. This paper thus studies the intensity-based color appearance of the image scene for IR applications, particularly in image browsing and searching. Databases or Digital Library systems that provide image-searching services always require an efficient method to deliver the best performance to customers. Therefore, a method built based on human judgment can help meet the essential needs and satisfaction of a human.

Any industry that has a large collection of big-sized digital color images actively provide browsing services to the customer (through the web or standalone) and will stand to benefit through this research. These industries could potentially be museums that deal with thousands of images or the entertainment industry as well as the design industries that deal with large collections of posters, designs, patterns, flyers, pamphlet etc. Ground-truth images determined from this work are samples that will be used as initial reference or guideline in our future research work, which will be focusing on image categorization based on color appearances in the area of IR.

## VII. LIMITATION AND FUTURE WORK

In acknowledging the limitations vested in this study, the various respondents who have participated in the survey were ones who were primarily available in accordance with the time the experiment was conducted. During this study, respondents were taken regardless of their age, work, gender of level of education. However, such respondents may not fully represent what could be accepted as the 'general public' or a universal symbol of humans on a daily basis.

To be fully considered as 'universally human', a certain extent must be achieved in which respondents are subject to complicated and deviating factors. Hence, to diversify the study, future work shall take into account the lacuna to fill in terms of respondent's careers, living conditions, fields of study and psychological factors. The future work will therefore supplement the results of this current study and catalogue each result based on a grander scale of groups of respondents.

In the future, for our work in IR application of landscape paintings, we plan to formulate a color appearance classifier based on the optimum color histogram we have obtained in our previous study. We will scrutinize the effectiveness

of the classifier using the Precision and Recall Method and evaluate the effectiveness by comparing with Support Vector Machines (SVM) and Naïve Bayes classifier. SVM and Naïve Bayes are amongst popular classification models [43].

## REFERENCES


[1] G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognit.*, vol. 46, no. 1, pp. 188–198, Jan. 2013, doi: 10.1016/j.patcog.2012.06.001.

[2] H. Farsi and S. Mohamadzadeh, "Colour and texture feature-based image retrieval by using Hadamard matrix in discrete wavelet transform," *IET Image Process.*, vol. 7, no. 3, pp. 212–218, Apr. 2013, doi: 10.1049/iet-ipr.2012.0203.

[3] B. G. Park, K. M. Lee, and S. U. Lee, "Color-based image retrieval using perceptually modified Hausdorff distance," *EURASIP J. Image Video Process.*, vol. 2008, no. 1, pp. 1–10, 2008, doi: 10.1155/2008/263071.

[4] N. Shrivastava and V. Tyagi, "An efficient technique for retrieval of color images in large databases," *Comput. Elect. Eng.*, vol. 46, pp. 314–327, 2015, doi: 10.1016/j.compeleceng.2014.11.009.

[5] J. Narwade and B. Kumar, "Local and global color histogram feature for color content-based image retrieval system," in *Proc. Int. Congr. Inf. Commun. Technol.*, 2016, pp. 293–300.

[6] K. Konstantinidis, I. Andreadis, and G. C. Sirakoulis, "Application of artificial intelligence methods to content-based image retrieval," *Adv. Imaging Electron Phys.*, vol. 169, pp. 99–145, Jan. 2011, doi: 10.1016/B978-0-12-385981-5.00003-3.

[7] K. Jang, S. Han, and I. Kim, "Person re-identification based on color histogram and spatial configuration of dominant color regions," 2014, *arXiv:1411.3410*. [Online]. Available: http://arxiv.org/abs/1411.3410

[8] M. A. Gavrielides, E. Sikudova, and I. Pitas, "Color-based descriptors for image fingerprinting," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 740–748, Aug. 2006, doi: 10.1109/TMM.2006.876290.

[9] S. G. Shaila and A. Vadivel, "Smooth weighted approach for colour histogram construction using human colour perception for CBIR applications," *Int. J. Multimedia Appl.*, vol. 4, no. 1, pp. 113–125, Feb. 2012, doi: 10.5121/ijma.2012.4110.

[10] S. Anwar Lashari, R. Ibrahim, N. S. A. Md Taujuddin, N. Senan, and S. Sari, "Thresholding and quantization algorithms for image compression techniques: A review," *Asia-Pacific J. Inf. Technol. Multimedia*, vol. 7, no. 1, pp. 83–89, Jun. 2018, doi: 10.17576/apjitm-2018-0701-07.

[11] L. Chen, "A new human perceptional color quantization algorithm," in *Proc. 3rd Int. Conf. Genetic Evol. Comput.*, Oct. 2009, pp. 710–713, doi: 10.1109/WGEC.2009.84.

[12] A. Othman, T. S. M. T. Wook, and S. M. Arif, "Quantization selection of colour histogram bins to categorize the colour appearance of landscape paintings for image retrieval," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 930–936, 2016, doi: 10.18517/ijaseit.6.6.1381.

[13] M. Ponti, T. S. Nazaré, and G. S. Thumé, "Image quantization as a dimensionality reduction procedure in color and texture feature extraction," *Neurocomputing*, vol. 173, pp. 385–396, Jan. 2016, doi: 10.1016/j.neucom.2015.04.114.

[14] C. Vasanthanayaki, "Color perception histogram for image retrieval using multiple similarity measures," *J. Comput. Sci.*, vol. 10, no. 6, pp. 985–994, Jun. 2014, doi: 10.3844/jcssp.2014.985.994.

[15] Y. Zhang, L. Gao, W. Gao, and J. Liu, "Combining color quantization with curvelet transform for image retrieval," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.*, Oct. 2010, pp. 474–479, doi: 10.1109/AICI.2010.105.

[16] D.-S. Park, T. Y. Kim, J. H. Han, and J.-S. Park, "Image indexing using weighted color histogram," in *Proc. 10th Int. Conf. Image Anal. Process.*, Sep. 1999, pp. 909–914.

[17] J. Huang, S. Kumar, W.-J. Zhu, R. Zabih, and M. Mitra, "Image indexing using color correlograms," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1997, pp. 762–768. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=609412

[18] M. Rundell, *Macmillan English Dictionary for Advanced Learners*. Oxford, U.K.: MacMillan Education, 2002.

[19] S. Wan, P. Jin, and L. Yue, "An effective image retrieval technique based on color perception," in *Proc. 6th Int. Conf. Image Graph.*, Aug. 2011, pp. 1017–1022, doi: 10.1109/ICIG.2011.165.

[20] M. Fairchild, "Color appearance terminology," in *Color Appearance Models*. Hoboken, NJ, USA: Wiley, 2013, pp. 85–96.

[21] S. Singh, "Impact of color on marketing," *Manage. Decis.*, vol. 44, no. 6, pp. 783–789, 2006, doi: 10.1108/00251740610673332.

[22] C. H. C. Leung and Y. Li, "Semantic image retrieval using collaborative indexing and filtering," in *Proc. IEEE/WIC/ACM Int. Conferences Web Intell. Intell. Agent Technol.*, Dec. 2012, pp. 261–264, doi: 10.1109/WI-IAT.2012.197.

[23] M. Gabbouj, M. Birinci, and S. Kiranyaz, "Perceptual color descriptor based on a spatial distribution model: Proximity histograms," in *Proc. Int. Conf. Multimedia Comput. Syst.*, Apr. 2009, pp. 144–149, doi: 10.1109/MMCS.2009.5256714.

[24] A. Othman, T. S. M. T. Wook, and S. M. Arif, "The analysis of colour appearance categories of landscape paintings for Malaysian users," in *Proc. 6th Int. Conf. Electr. Eng. Informat. (ICEEI)*, Nov. 2017, pp. 1–6, doi: 10.1109/ICEEI.2017.8312397.

[25] A. Othman and K. Martinez, "Colour appearance descriptors for image browsing and retrieval," *Proc. SPIE*, vol. 6820, Jan. 2008, Art. no. 68200R, doi: 10.1117/12.766882.

[26] C. Rigaud, D. Karatzas, J.-C. Burie, and J.-M. Ogier, "Color descriptor for content-based drawing retrieval," in *Proc. 11th IAPR Int. Workshop Document Anal. Syst.*, Apr. 2014, pp. 267–271, doi: 10.1109/DAS.2014.70.

[27] R. Ashraf, M. Ahmed, S. Jabbar, S. Khalid, A. Ahmad, S. Din, and G. Jeon, "Content based image retrieval by using color descriptor and discrete wavelet transform," *J. Med. Syst.*, vol. 42, no. 3, p. 44, Mar. 2018, doi: 10.1007/s10916-017-0880-7.

[28] E. C. Kee and G. Lawson, "A psychophysical study of colour perception using digital and real objects," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 59, no. 1, pp. 1307–1311, Sep. 2015, doi: 10.1177/1541931215591214.

[29] A. Mojsilovic, "A computational model for color naming and describing color composition of images," *IEEE Trans. Image Process.*, vol. 14, no. 5, pp. 690–699, 2005, doi: 10.1109/TIP.2004.841201.

[30] A. Mojsilovi, J. Gomes, and B. Rogowitz, "ISee: Perceptual features for image library navigation," *Proc. SPIE*, vol. 4662, pp. 266–277, May 2002, doi: 10.1117/12.469523.

[31] M. H. Kim, T. Weyrich, and J. Kautz, "Modeling human color perception under extended luminance levels," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 1–9, Jul. 2009, doi: 10.1145/1531326.1531333.

[32] J. Lee and E. Park, "Fuzzy similarity-based emotional classification of color images," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1031–1039, Oct. 2011, doi: 10.1109/TMM.2011.2158530.

[33] J. Van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1512–1523, Jul. 2009, doi: 10.1109/TIP.2009.2019809.

[34] S. Pattanaik and D. G. Bhalke, "Efficient content based image retrieval system using Mpeg-7 features," *Int. J. Comput. Appl.*, vol. 53, no. 5, pp. 19–24, Sep. 2012.

[35] D. Neumann and K. R. Gegenfurtner, "Image retrieval and perceptual similarity," *ACM Trans. Appl. Perception*, vol. 3, no. 1, pp. 31–47, Jan. 2006, doi: 10.1145/1119766.1119769.

[36] J. A. Lay and L. Guan, "Retrieval for color artistry concepts," *IEEE Trans. Image Process.*, vol. 13, no. 3, pp. 326–339, Mar. 2004.

[37] M. Tai, R. Y.-H. Yang, and S. Liao, "A study of color description and cognition," in *Proc. 3rd Int. Conf. Inf. Sci. Interact. Sci.*, Jun. 2010, pp. 138–141, doi: 10.1109/ICICIS.2010.5534726.

[38] C. N. W. Giragama, C. A. Marasinghe, A. P. Madurapperuma, D. R. Wanasinghe, S. Herath, and O. Minetada, "Cross-language similarity between perceptual and semantic structures of color tones," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, vol. 1, Oct. 2006, pp. 345–352, doi: 10.1109/ICSMC.2006.384406.

[39] T. S. M. Tengku Wook, H. Mohamed, N. S. Ashaari, S. F. Mat Noor, Z. Muda, I. Y. Zairon, and F. L. Khaleel, "User experience evaluation towards interface design of digital footprint awareness application," *Asia-Pacific J. Inf. Technol. Multimedia*, vol. 9, no. 1, pp. 17–27, Jun. 2020, doi: 10.17576/apjitm-2020-0901-02.

[40] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *Brit. J. Math. Stat. Psychol.*, vol. 61, no. 1, pp. 29–48, 2008, doi: 10.1348/000711006X126600.

[41] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.

[42] S. T. Wong, N. Saddki, and W. N. Arifin, "Inter-rater reliability of the Bahasa Malaysia version of patient education materials assessment tool," *Med. J. Malaysia*, vol. 74, p. 100, Aug. 2019.

[43] W. D. Ahmad and A. A. Bakar, "Classification models for higher learning scholarship award decisions," *Asia-Pacific J. Inf. Technol. Multimedia*, vol. 7, no. 2, pp. 131–145, Dec. 2018, doi: 10.17576/apjitm-2018-0702-10.

**ANIZA OTHMAN** received the degree (Hons.) in computer science from Universiti Putra Malaysia (UPM), in 1993, and the master's degree in computer science (multimedia), in 2000. She is currently pursuing the Ph.D. degree with Universiti Kebangsaan Malaysia. She worked as a Research Assistant with the Engineering Faculty and Electronics, Universiti Putra Malaysia, where she had the opportunity to get involved in developing programs for robotic (RT100) using PASCAL. In 2001, she became a Lecturer with University Teknikal Malaysia Melaka. She was attached to the Department of Interactive Media, Faculty of Information and Communications Technology, before joining the Department of Software Engineering of the same Faculty. Her research include interest the areas of color appearance features, color image analysis and classification, color image browsing and searching, and teaching/learning applications and technologies.

**TENGKU SITI MERIAM TENGKU WOOK** received the bachelor's and master's degrees in information technology from the National University of Malaysia, in 1998 and 1999, respectively, and the Ph.D. degree in human–computer interaction from University Malaya, Malaysia, in 2012. She is currently a Senior Lecturer with the Software Technology and Management Research Center and also the Head of Masters Programme Studies at the Faculty of Information Science and Technology, National University of Malaysia. She has produced more than 20 indexed journals (ISI / Scopus). Meanwhile, there are 25 ISI and Scopus indexed proceedings, as well as 10 other indexed proceedings. In addition, she has also produced five chapters in book. She is currently writing a book about virtual heritage as well as leading a Prototype Research Grant Scheme (PRGS) on expertise User Interaction. She has received over 14 external and internal research grants such as Fundamental Research Grant Scheme (FRGS), University Grant (GUP), Top Down and Innovation Grant, with a total of more than 1 million. Her research focuses on the human–computer interaction that includes the multimedia application, user interaction design and usability, virtual and augmented reality, and also e-learning.

**FAIZAN QAMAR** received the B.E. degree in electronics from Hamdard University, Karachi, Pakistan, in 2010, the M.E. degree in telecommunication from NED University, Karachi, in 2013, and the Ph.D. degree in wireless networks from the Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia, in October 2019. He is currently serving as a Senior Lecturer with the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Selangor, Malaysia. He has more than eight years of research and teaching experience. He has authored and coauthored numerous ISI & Scopus journals and IEEE conference papers. His research interests include interference management, millimeter-wave communication, the IoT networks, D2D communication, and Quality of Service enhancement for future wireless networks. He is also a reviewer of several national and international journals and the IEEE conferences proceedings.