"I Have No Text in My Post": Using Visual Hints to Model User Emotions in Social Media

Junho Song
junonjuno@bigdas.hanyang.ac.kr
Department of Computer Science, Hanyang University
Seoul, South Korea

Kyungsik Han*
Kyungsikhan@hanyang.ac.kr
Department of Intelligence Computing, Hanyang University
Seoul, South Korea

Sang-Wook Kim*
wook@hanyang.ac.kr
Department of Computer Science, Hanyang University
Seoul, South Korea

ABSTRACT

As an emotion plays an important role in people's everyday lives and is often mirrored in their social media use, extensive research has been conducted to characterize and model emotions from social media data. However, prior research has not sufficiently considered trends of social media use—the increasing use of images and the decreasing use of text—nor identified the features of images in social media that are likely to be different from those in non-social media. Our study aims to fill this gap by (1) considering the notion of *visual hints* that depict contextual information of images, (2) presenting their characteristics in positive or negative emotions, and (3) demonstrating their effectiveness in emotion prediction modeling through an in-depth analysis of their relationship with the text in the same posts. The results of our experiments showed that our visual hint-based model achieved 20% improvement in emotion prediction, compared with the baseline. In particular, the performance of our model was comparable with that of the text-based model, highlighting not only a strong relationship between visual hints of the image and emotion, but also the potential of using only images for emotion prediction which well reflects current and future trends of social media use.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Emotion analysis, social media images, visual hints

ACM Reference Format:
Junho Song, Kyungsik Han, and Sang-Wook Kim. 2022. "I Have No Text in My Post": Using Visual Hints to Model User Emotions in Social Media. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485447.3512009

1 INTRODUCTION

An emotion is an indicator of a mental state and plays an important role in one's daily life, including social interactions, self-reflection, and decision-making [13]. As social media has become an important source for expressing and sharing one's emotions, a great amount of data embedded with emotions are generated online. This also accelerates research interests and efforts in various domains, ranging from psychology to social science, investigating characteristics of emotions and developing computational models that predict emotions [4, 8, 18–20, 39, 41, 43]. In particular, research in natural language processing has focused on the development of emotion prediction models using features from language information [7, 30, 31, 36], and that in computer vision has modeled emotions through image information [9, 14, 33].

Most social media platforms support social interactions through various types of information, such as images, videos, and text. While the role and affordance of each type vary, people are now exhibiting strong preferences for social media platforms that support image- or video-oriented interactions. According to recent reports [26, 28, 29], the usage of image- and video-based platforms (e.g., Tiktok, Instagram) tends to increase, whereas that of text-based platforms (e.g., Twitter) shows the opposite. Similar perspectives are also observed in our dataset (discussed in Section 4.1).

The increasing use of such media-based social platforms suggests that it may be necessary to handle the data in those platforms from a different perspective, not the same as that of the existing text-centered emotion studies [6, 7, 27, 30, 31, 36–38]. The text in social media posts becomes shortened or often contains a simple list of hashtags rather than complete sentences. Having less text in the data is likely to degrade the performance of models whose dependency on text is high. To overcome this, some recent studies pay attention to the images of the posts [34, 43, 45]. They utilize the raw features (e.g., colors, textures, layout, balance) of the images as a supplement for text information. However, they still lack the consideration of *contextual information* and fail to provide reasons and interpretations of how the model algorithms operate. The utilization of images in such ways may not give sufficient information that relates to user emotion or the context that reflects the emotion. A recent survey on affective research also emphasized the necessity of modeling *image context* for image analysis [45].

Therefore, we aim to address a shortage and limited use of image information. We utilize the *features that correspond to contextual information* of images, namely *visual hints*. Our work begins with the confirmation of our research motivations. We deal with the characterization of visual hints, validation of the effects of visual hints on emotion prediction, and discussion of study implications. In this paper, we aim at answering the following research questions:

- RQ1. How do users express their emotions in the posts?
- RQ2. How do visual hints relate to emotions?
- RQ3. How valid are visual hints for predicting emotions?

*Corresponding authors.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512009

To answer RQ1, we conducted an online survey that inquires into ways of presenting one's emotions on Instagram posts to 300 real Instagram users through Amazon Mechanical Turk. As a result, 61.4% of the respondents who chose the images answered that they tend to express their emotions through *visual hints*, rather than color, brightness, or texture in the image, which confirmed our motivation.

To answer RQ2, we crawled 108,108 public posts on Instagram and analyzed the images of the posts. The results indicated that visual hints were clearly different between the two emotions, indicating that visual hints can be used as the important features for emotion modeling.

To answer RQ3, we built emotion prediction models and confirmed the effectiveness of visual hints on model performance. Our model yielded 18-24% higher performance (in f1-score) than did the models only with raw image information (i.e., color) and 6-12% higher performance than the recent benchmarks (i.e., fine-tuned CNN models, ResNet50, VGG16, AlexNet, and Inception_V3), respectively. Also, our model showed comparable performance to the model with text information, highlighting a salient role of visual hints on presenting emotions and reflecting current trends in social media use.

Based on the results, we confirmed the strong relationship between visual hints and emotions in social media posts. We also validated the effectiveness and robustness of visual hints as the features for emotion modeling, even with the posts that have little text information, which reflects current trends of social media use.

2 RELATED WORK

Emotion is a generic term. Since the emotions each person feels are subjective even in the same situation, it is uneasy to specify an emotion by physical expression, biological reaction, mental state, and so on [13]. Accordingly, much research has analyzed the relationship between people's emotions and expression methods. As recent social media platforms provide users with an opportunity to post their content with various media types (e.g., images, videos, text), user generated content has become richer and more diverse, so do ways of emotion expression in the content.

2.1 Understanding emotions in images

Recent research on emotions through image information has been proceeding in the direction of extracting various features of images and applying those features to model construction, thereby confirming its effectiveness. According to Zhao et al. [45], those features can be categorized into three levels (i.e., low, mid, and high).

First, low-level features include color or texture, which somewhat lack reasonable interpretation. Color is one of the most basic raw features of an image. Valdez and Mehrabian [34] studied the relationship between color and human emotions, and found that saturation and brightness had stronger relationships with emotions than did hue. Babin et al. [1] studied how color and light affect people's emotion about shopping items and verified changes of the participants' emotions by the color and light intensity of items.

Second, mid-level features are more human-interpretable than low level features. For example, Zhao et al. [43] employed the Principle-of-Art-based Emotion Feature (PAEF), characterized by a combination of balance, emphasis, harmony, variety, gradation, and movement from the images. The authors extracted PAEF features from the artistic photograph and the abstract painting datasets and developed a model that classifies eight emotions, yielding a range of 0.55–0.72 precisions per emotion. Patterson and Hays [22] presented a taxonomy of 102 discriminative attributes, including materials (e.g., cement), surface properties (e.g., rusty), functions (e.g., cooking), spatial envelop attributes (e.g., symmetric), and object presence (e.g., chairs). With this dataset, Yuan et al. [42] presented an emotion classification model that achieved 0.64 accuracy.

Third, high-level features present the semantic information of images, which are more interpretable and can be easily understood by humans. Studies have considered facial expressions [40], size of faces [20], and semantic concepts based on pairs of adjective noun words (e.g., beautiful flower, crying baby) [5]. There are two datasets developed based on the idea of adjective and noun pairs, SentiBank [5] and MVSO [16], and the performance of the emotion classification models based on these two datasets ranged from 0.30 to 0.77 accuracies, varied by test datasets [45].

Recent research has used deep features by employing machine/deep learning. You et al. [41] evaluated a CNN-based model that classifies seven emotions from images on Flickr and Instagram. They trained the model based on the pre-trained ImageNet-CNN, achieving 0.58 accuracy. With images collected from Flickr, Zhao et al. [44] proposed an emotion prediction model based on rolled multi-task hypergraph learning, considering visual content (object), social context (user relationships), temporal evolution (emotion sequence), and location influence (image meta data). The model showed averages of 0.49 precision, 0.09 recall, and 0.33 f1-score.

2.2 Understanding emotions in text

Using keywords and sentences is the most basic way to identify text features. Theresa et al. [37] extracted syntactic and subjective clues from the literature and used as features to classify every clause of a sentence according to the strength of emotion. Wu, Chuang, and Lin [38] utilized the semantics from a sentence and found their associations with emotions. Bracewell [6] suggested keyword-based emotion detection, which is based on the semi-automatic construction of a keyword dictionary. The keyword dictionary collects both the keywords and the relations among them. Su et al. [27] uncovered the relationship between words and sentiments.

Compared to images, relatively more efforts have been exerted in extracting and utilizing contextual information from the text of posts in emotion classification research. Emotion modeling with text information [6, 27, 30, 31, 36–38] has been elaborated through a word embedding method [11, 21, 24] that utilizes not only words but also context information, such as order and relationship between words. Cai and Hofmann [7] combined information-theoretic measures and semantic knowledge of a hierarchy using WordNet [32] to extract concepts from texts automatically. Turney [30, 31] and Wang et al. [36] used a bag of words and semantic information to enrich the representation of text classification.

2.3 Emotion modeling from social media posts

Our summary of literature review highlights the following two perspectives.

First, existing research has mainly focused on the inherent (low- and mid-level features) or text dependent (high-level features) characteristics of images. Some research used conventional images (e.g., paintings), where general perspectives were embedded (e.g., bright

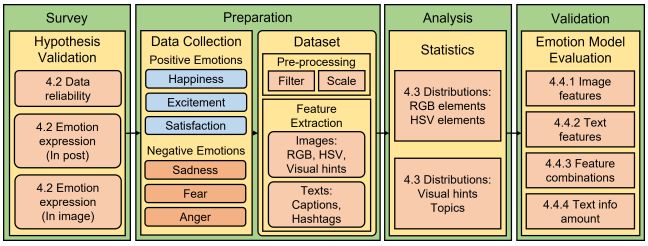


Figure 1: A study procedure and methodology.

Figure 2: An example of visual hints from an image.

color means positive emotion, dark color means negative emotion), not actually from those that were produced and shared on social media [34, 43]. Deep learning-based modeling still lacks transparent explanation about the relationship between one's emotion and image characteristics. Second, methods applied in prior text-based studies can operate *only when the content has sufficient text information*. As social media posts have become diversified due to the high use of other media types (e.g., image, video), emotion modeling with text only has become less effective nor sustainable. Information diversity is one of the key aspects that characterizes social media [10]. The contextual information of images often includes a user's intentions (one of which is emotion sharing) [9, 14, 33], yet has not been sufficiently considered in understanding emotions [45].

Thus, in this paper, we present the role of *visual hints* in images on emotion classification. Specifically, we identify characteristics of image information, measure their relationships with emotions, examine a possibility of classifying emotions without relying on text information. Our work reflects the trend of social media usage (reduction in text usage) and further discusses the utilization of the visual hints in emotion classification.

3 STUDY PROCEDURE AND METHODOLOGY

Figure 1 shows our study procedure, consisting of four phases: First, we verify our hypothesis through a survey that asks real Instagram users about how visual hints are used in expressing emotions (hypothesis verification). Second, we collect public posts on Instagram and extract image and text information from each post (preparation). Third, we conduct an analysis of image information for each of the positive and negative emotions and confirm the relationship between each emotion and visual hints (analysis). Finally, we validate the visual hints through the evaluation of an emotion classification model (model validation).

3.1 Survey

We conducted a survey with 300 real Instagram users through Amazon Mechanical Turk. We asked the participants to indicate their Instagram page link for a user verification purpose[1]. We used the answers only from those whose Instagram page is valid. The survey consisted of the following three survey questions (SQ):

- SQ1. Do emotion hashtags in your posts honestly express your emotions?
- SQ2. How do you express your emotions when adding posts on Instagram? (multiple choice question)
- SQ3. When posting a post on Instagram, how do you express your emotions in images? (multiple choice question)

Our intention for each question was as follows. SQ1 was to verify whether using a hashtag is appropriate for data collection and *emotion labeling*. SQ2 was to check whether people show their emotions in the image. SQ3 was to examine how people's emotions are expressed in the image, which is a more specific version of SQ2. The survey took about two minutes, and each respondent was paid with $0.5 for their time.

[1]We considered users who are active on Instagram (having a post within the most recent six months).

3.2 Preparation

3.2.1 Data collection. We collected public posts from Instagram in which all posts have image content. As one of the most popular social media platforms, Instagram well represents current trends in social media use. To this end, we employed six emotions (happiness, excitement, and satisfaction as positive; sadness, anger, and fear as negative) that have been frequently used in many prior studies [5, 8, 12, 18–20, 23, 44]. We used the hashtag of each emotion (#happy, #excite, #satisfy, #sad, #angry, and #scary) and searched the posts by those hashtags. We collected more than 20,000 posts for each hashtag through the Graph API (i.e., official open API for developers provided by Instagram)[2]. We excluded spam posts (e.g., advertisements, bots), the posts with different proportions of images (e.g., posts that are not square images), and the posts containing two or more emotions (e.g., the posts having two or more emotion hashtags used for collection). To this end, for every collected posts, two authors of this work further reviewed the remaining posts and emotions, and only used the posts that both authors agreed with for modeling and analysis. As a result, we used 108,108 posts (positive: 56,285, negative: 51,823) for the analysis.

3.2.2 Feature extraction. From the collected posts, we extracted the image information (i.e., colors and visual hints) and the text information (i.e., captions and hashtags). Here, the size of the image was re-adjusted to 150×150 to increase the efficiency of data analysis and model evaluation. Although Instagram supports various sizes of images, we used only the square images (1:1) because they are the most basic and widely used. Images of different ratios were excluded because bias may occur in the process of transforming the images at the same ratio (i.e., adjusting the ratio of images may distort the distribution of features), and this will affect statistical analysis.

We extracted the color information of the image per pixel level based on the RGB model that expresses color information as a combination of Red, Green, and Blue elements, and the HSV model that expresses the combination of Hue, Saturation, and Value elements. Next, the number of pixels of each element was converted into a vector, which represents the distribution of color elements per image. In the case of the RGB model, 768 elements were converted into a vector with the number of pixels in the image corresponding

[2]We used the "research_ap", one of the permissions, which allows developers to collect public Instagram posts. In March 2021, Instagram released a new Analytics API (https://research.fb.com/blog/2021/03/new-analytics-api-for-researchers-studying-facebook-page-data/); the Graph API has been deprecated since then.

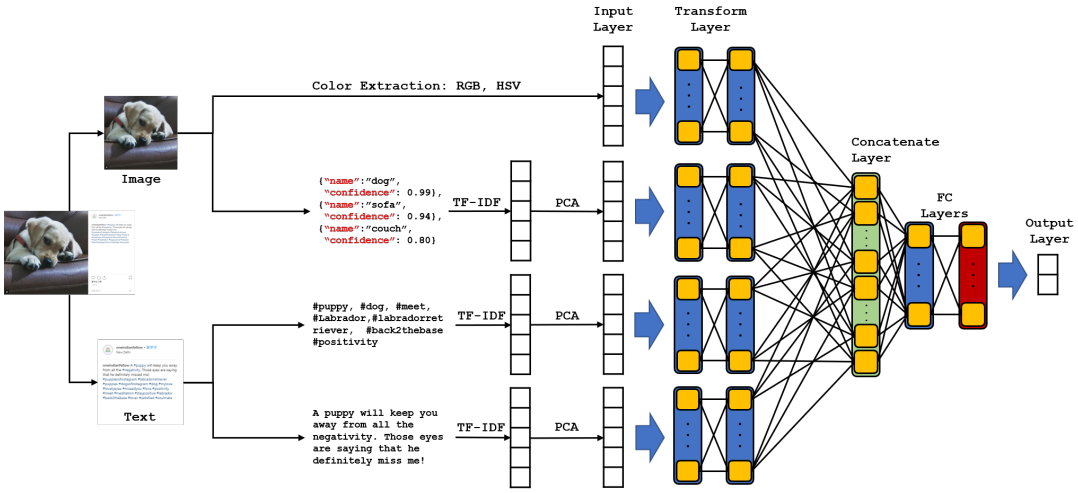


Figure 3: The emotion classification model based on a fully connected network (for all features).

to r0 to r255, g0 to g255, and b0 to b255 (in the case of HSV model, 560 elements were converted into h0-h359, s0-s99, v0-v99).

For obtaining visual hints, we used the MS Azure Cognitive Service API, which extracts not only the objects of the image, but also the action and state information derived from some of the objects. The API has shown its effectiveness in many practical applications, showing its validity for use. For example, from the image shown in Figure 2, the main object of the woman and other objects, such as sky, building, bridge, and sand, were extracted. Additional information, such as looking and clear, which indicate the status of the objects, were also extracted. Each aspect of the extracted information has a *confidence value* likely to lead to classification accuracy. We used the information with a confidence value over 0.7 as a visual hint to ensure the accuracy.

For text information, we extracted the hashtags and captions from each post. In the case of a hashtag, it is in the form of a "#" symbol and a word, thus, we used the words excluding #. In the case of a caption, it has the form of a short sentence; thus, we excluded stopwords after tokenization. We also excluded the six emotion words to make sure only non-emotional words were used in the analysis and modeling.

Visual hints and text information were converted to numeric values by a TF-IDF [25] transformer trained with the entire dataset. Since the visual hint for an image is an arrangement of words, we decided that TF-IDF, which is the context-free method, is appropriate. We also considered other embedding methods, Word2Vec (which learns the order of words in the sentence) [21] and Doc2Vec (which learns the context within the document) [17]. However, they did not show significant differences in the performance of emotion classification models in our preliminary experiments. We thus used TF-IDF for modeling.

3.3 Data analysis

We analyzed the colors and the visual hints of the images that were extracted from our dataset in the preparation phase. First, we checked the color distribution of images corresponding to the positive and negative emotions. The color distribution was confirmed through the RGB elements; the color saturation and brightness distribution were confirmed through the HSV elements.

Table 1: The settings of emotion classification models.

	Layers	Activation (hidden)	Pooling	Dropout (rate)	Activation (output)	Loss
FCN	6 dense	ReLU	-	6 layers (0.5)	Sigmoid	MSE
CNN	3 conv + 6 dense	ReLU	3 layers (max)	9 layers (0.5)	Sigmoid	MSE

Next, we checked the distribution of the visual hint information. We classified the visual hints by topics, using Latent Dirichlet Allocation (LDA) [3], one of the topic modeling techniques. LDA requires specification of the number of topics. Considering a bottom-up approach, we started from 100 topics, manually examined and merged similar ones. Since each LDA run returns slightly different results, we conducted this process until we reached the consensus on the topics. As a result, we identified 15 topics as follows: `state`, `activity`, `animal`, `anniversary`, `color`, `entertainment`, `fashion`, `food`, `indoor`, `outdoor`, `nature`, `person`, `object`, `text`, `urban`. The visual hints of each topic were also identified. Based on this, we investigated the distributions of the topics and their visual hints by emotions.

3.4 Emotion classification model

We evaluated the effectiveness of visual hints as a feature for emotion modeling. To this end, the performance of the emotion classification model with visual hints based on machine and deep learning and that with other features were compared. We built emotion classification models based on SVM, Logistic Regression, Random Forest, and Decision Tree. In the case of deep learning models, we employed a fully connected network (FCN) model. We also benchmarked not only a convolutional neural network (CNN) [35] but also transfer learning with computer vision models (i.e., AlexNet, VGG19, ResNet50, and InceptionV3) which all have shown great performance in many recent studies [34, 43, 45]. Regarding the features, we used color (RGB/HSV) distributions, visual hints, hashtags, and captions.

The FCN-based emotion classification model consists of four parts of layers (Figure 3): input layers having the same dimensionality as feature vectors, transform layers that convert each of inputs' dimensionality to the same dimensionality (500), a concatenate

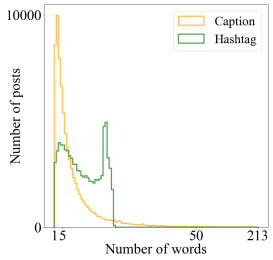


Figure 4: The distribution of posts on the number of words.

layer that combines the features, and fully connected layers that return the probability of the positive and negative emotions. In the case of the CNN-based emotion classification model, a 150×150 image with pixel values on R, G, and B channels was used as input. The model consists of convolutional layers that perform convolution with filters of 3×3, 5×5, and 9×9 and fully connected layers that return the probability of positive and negative emotions. Mean Squared Error (MSE) is employed as the loss function for all models. For transfer learning models, we used the pre-trained vision models (i.e., AlexNet, VGG19, ResNet50, and InceptionV3) and re-trained an output layer with our dataset to fit our emotion classification purpose. The input size of the image is the same as the one used in the CNN-based model. Detailed settings for each model are summarized in Table 1.

Finally, 80% of the data was used as a training set and 20% as a test set. For the emotion classification model based on deep learning, 10% of the training set was used as a validation set at each epoch. We performed optimization using the Adam optimizer for up to 100 epochs, and an early stop strategy was used with 20 times of patience. The best parameters of all models were selected through parameter-tuning with the validation set. Each emotion classification model performs 5-fold cross-validation.

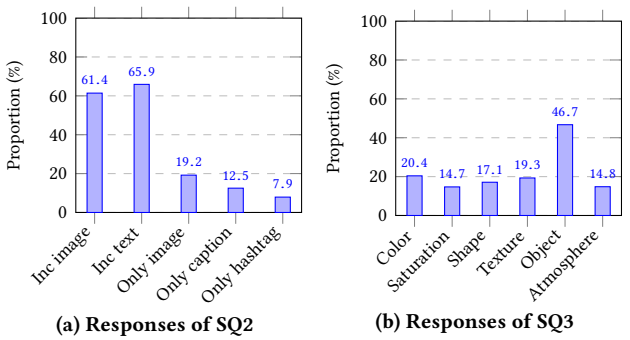
4 RESULTS

4.1 Sparseness of text information

As highlighted in the introduction section, we observed similar patterns in our dataset (108,108 posts). The percentage of the posts with no text (i.e., union of hashtags and captions) is 15%. The median number of the words in captions was only 5 (min: 0, max: 207), and that in hashtags was 13 (min: 0, max: 31). As illustrated in Figure 4, the graphs of captions and hashtags are significantly skewed, such that the large portion of the posts has a small amount of text. This phenomenon is more prominent in captions than in hashtags. These results indicate the sparseness of text information in image-based social media posts.

4.2 Emotion labels

As a result of reliability in expressing emotions through hashtags (SQ1), 74% of participants answered that the *emotion hashtags* they wrote in posts correspond to their emotions. Only 6% answered that the hashtags do not relate to their emotions, and 20% said it depends on the case. These results are consistent with the results of existing research that has found that hashtags are used to express user intentions or emotions [15]. Therefore, our rationale of using the emotion hashtags for data collection is well supported, and each emotion hashtag can be used as the label of each post (i.e., ground truth).



(a) Responses of SQ2 **(b) Responses of SQ3**

Figure 5: User responses to ways of expressing emotions (a) in social media posts and (b) in the images.

4.3 RQ1. Emotions expression in posts

As a result of ways of expressing emotions in social media posts (SQ2, multiple choices allowed, Figure 5a), 61.4% of participants answered that they expressed their emotions *through images*, and 65.9% of participants answered that they expressed their emotions through text. Among the responses that received only a single choice, images were chosen by 19.2% of participants, higher than captions (12.5%) and hashtags (7.9%). From these results, we confirmed that the use of images is highly frequent as a way to express one's emotions in the posts.

As a result of ways to express emotions in the images of social media posts (SQ3, allowed multiple choices, Figure 5b), 46.7% of respondents answered that they expressed emotions through the objects in images. The raw features of images, such as color, saturation/brightness, shape, and texture, which were much considered in prior studies, were selected much less than the objects (even the color was less than half of the object).

In summary, these results that reflect end-users' perspectives and experiences in the context of social media, well confirmed that the "visual hints" of the image are valid information by which to express the user's feelings in social media posts.

4.4 RQ2. Visual hints and emotions

First, the color distributions in the positive and negative emotions did not show significant differences ($p > 0.05$) for both the RGB and the HSV elements. Interestingly, this result is *not consistent* with the one presented in prior studies [1, 34]. We will discuss such different results later in Section 5.1.

Table 2 summarizes examples of visual hints in the positive and negative emotions by topic. It is worth noting that visual hints are quite different by emotion. Among them, differences of visual hints between the positive and negative emotions were remarkable in the topics of `state`, `animal`, `color`, `fashion`, `food`, and `object`.

Figure 6 illustrates the distribution of the topics for each emotion. The frequent topics in the images of positive emotions are `animal`, `fashion`, or `person`, while those of negative emotions are `state`, `color`, `person`, or `object`. Both emotions had `person` as the most frequently appearing topic, because of the nature of social media posts: sharing one's daily life. Among the positive emotions, the topics related to `fashion` were shown the most frequently in happiness and excitement, and those related to `food` were shown the most frequently in satisfaction. On the contrary, in the negative emotions, `object`, `color`, and `status` appeared the most frequent topics in sad, scared, and angry, respectively.

Table 2: Examples of visual hints in the positive (P) and negative (N) emotions for each topic (note that anniversary shows only two visual hints in the negative emotion).

Topic		Visual hints				
state	P	decorated	aquatic	blur	finned	progress
	N	sink	stationary	old	stuffed	wild
activity	P	travel	sport	pool	fishing	skiing
	N	skating	tennis	golf	boxing	skiing
animal	P	pet	turtle	cat	bird	dog
	N	reptile	rhinoceros	gull	lion	lizard
anniversary	P	birthday	decoration	party	ceremony	wedding
	N	birthday	wedding	-	-	-
color	P	white	red	blue	pink	green
	N	dark	red	blue	grey	black
entertainment	P	book	stage	screen	music	game
	N	puzzle	stage	gallery	resort	game
fashion	P	shoes	hat	costume	jacket	accessory
	N	scarf	cosmetic	dress	tattoo	mask
food	P	cake	bowl	drink	fruit	ice cream
	N	donut	coffee	pizza	chocolate	snack
indoor	P	kitchen	dressroom	ceiling	window	floor
	N	kitchen	toilet	bathroom	dish	door
nature	P	mountain	sunset	sky	grass	river
	N	valley	cliff	wave	lake	canyon
outdoor	P	road	park	bench	ground	pier
	N	wall	park	field	bridge	gravestone
person	P	man	posing	prey	head	sitting
	N	head	eye	girl	boy	man
object	P	toy	brick	clock	gift	goods
	N	wire	toiletry	weapon	doll	candle
text	P	newspaper	sign	map	board	recipe
	N	newspaper	sign	book	board	graffiti
urban	P	city	car	building	street	transport
	N	sidewalk	transport	motorcycle	hospital	bus

Figure 7 shows top-20 most frequently appearing visual hints. The images that have a person were the most frequent in both emotions; that is, it is consistent with the topic of frequent visual hints shown in Figure 6 and might be due to the nature of social media. In addition, the frequencies of the visual hints between the positive and negative emotions were different. Some of the visual hints, such as text, man, animal, and black, were more frequent in the negative emotion, whereas others, such as food, plate, and table, were more frequent in the positive emotion. These differences indicate that each visual hint can be a feature that represents each emotion.

4.5 RQ3. Validity of visual hints

We evaluated emotion classification models to verify the validity of visual hints in emotion classification. Table 3 summarizes the performances of the emotion classification models over different features. To examine the influence of text and image features on emotion classification, we present the results of model performance through the lens of features, word representation, and the amount of text information.

4.5.1 Image features. First, when the color features (RGB and HSV rows) were used, the models based on machine learning (SVM, Logistic Regression, Random Forest, and Decision Tree) and those based on deep learning (FCN) yielded f1-scores between 0.55-0.60 and between 0.57-0.59, respectively. On the other hand, with the visual hint features, the performance of the models based on machine learning and that based on deep learning increased to 0.63 and 0.71 f1-scores, respectively, which are 7% and 20% improvement, compared to the model with the color features.

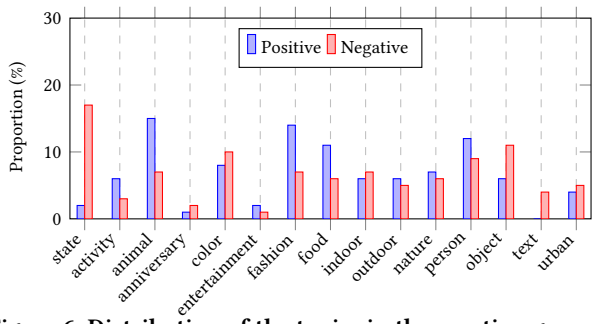


Figure 6: Distribution of the topics in the emotion groups.

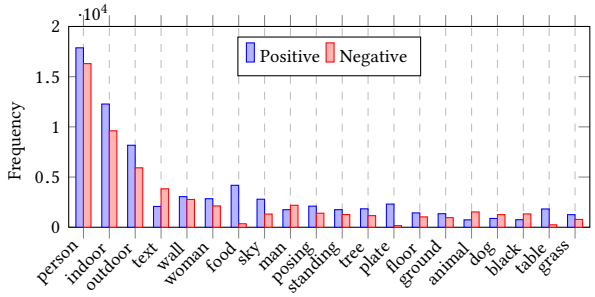


Figure 7: Top-20 most frequently appearing visual hints. The images with person, indoor, and outdoor appeared the most frequently in both emotions, which is related to the use of social media for sharing purposes.

Table 3: Performance comparisons of the emotion classification models (f1-score). The model with the visual hint features showed 20% better performance than that with the color features and showed a comparable performance (2-3% differences) to that with the text features.

Features	SVM	LR	RF	DT	FCN
RGB	0.58	0.58	0.55	0.58	0.57
HSV	0.60	0.60	0.57	0.60	0.59
Visual hint	0.62	0.63	0.62	0.62	0.71
Hashtag	0.70	0.70	0.70	0.69	0.74
Caption	0.70	0.70	0.70	0.69	0.73
Color (RGB+HSV)	0.59	0.60	0.56	0.60	0.57
Image (RGB+HSV+Visual hint)	0.64	0.64	0.60	0.64	0.71
Text (Hashtag+Caption)	0.70	0.70	0.70	0.66	0.74
All	0.74	0.74	0.70	0.73	0.76

4.5.2 Text features. Second, when the hashtag features were used, the models based on machine learning and deep learning yielded about 0.69-0.70 and 0.74 f1-scores, respectively. With the caption features, the models based on machine learning and deep learning yielded 0.69-0.70 and 0.73 f1-scores, respectively. These results are consistent with those in prior studies (i.e., text information has a

strong correlation with emotion), which was confirmed again in our study. When comparing the performances of the FCN model based on the hashtags or captions with those based on the visual hints, the difference was small (2-3%). This indicates comparable performance of the model with visual hints only. As highlighted in our research motivations, the image always exists but not the text in the post; thus the utilization of the image for emotion modeling becomes important. This result of comparable performance highlights a high potential of the visual hint features for emotion classification.

4.5.3 Feature combinations. Third, Table 3 summarizes the model performance according to the combination of features. In the case of image features using the RGB, HSV, and visual hints together, the FCN-based model yielded a 0.71 f1-score. The model with color and visual hints did not show performance improvement over that with visual hints only (both are 0.71 f1-scores). This indicates little influence of the color features on model performance. Similarly, there was no performance improvement from the model with hashtags only (0.74 f1-score) to those with hashtags and captions (0.74 f1-score). Finally, when all features were used, the FCN-based model yielded 0.76 f1-score, showing a 7% performance improvement compared with the model based on the image features (0.71 f1-score) and 2% improvement compared with the model based on the text features (0.74 f1-score).

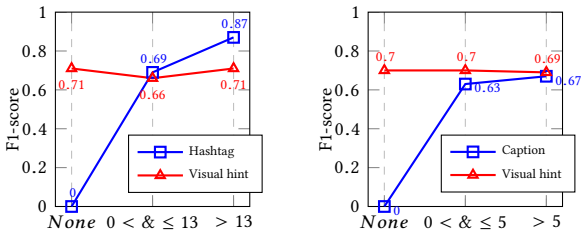
Table 4: Performance of the emotion classification models based on fine-tuning and transfer learning. Our model with visual hint features showed the best performance.

Models	Accracy	Precision	Recall	F1-score
Fine-tuned CNN	0.52	0.51	0.92	0.63
ResNet	0.52	0.51	0.96	0.67
VGG	0.53	0.52	0.78	0.63
AlexNet	0.54	0.52	0.84	0.64
Inception	0.52	0.51	0.80	0.63
Visual hint FCN (Ours)	0.61	0.58	0.92	**0.71**

We further compared our model with other image-oriented models based on fine-tuning and transfer learning (e.g., Fine-tuned CNN, ResNet, VGG16, AlexNet, Inception_V3) that were discussed in the recent survey paper [45]. As shown in Table 4, our model yielded the highest performance, again highlighting the effectiveness and robustness of visual hints in emotion modeling.

4.5.4 Text information amount. Lastly, we measured the reliability of the model depending on the availability of data of image and text. Figure 8 illustrates the model performance according to the amount of text information. We grouped the posts based on the *median* number of the words in hashtags and had three groups as follows: [no hashtags, 0<hashtags≤13, hashtags>13] (Figure 8a). We applied the same method for the captions and had three groups as follows: [no captions, 0<captions≤5, captions>5] (Figure 8b). From the posts in each group by hashtags, we built the FCN-model using hashtags and the one using visual hints. We did the same from the posts in each group by captions.

As a result, the performance decreased as did the amount of text information. For the hashtags with the FCN-based model, the group with the fewest (0<hashtags≤13) and that with the most



(a) Hashtag vs Visual hint **(b) Caption vs Visual hint**

Figure 8: Model performance according to the amount of text information. The performances of the model with visual hints were quite steady, while those of the model with text information significantly varied.

(hashtags>13) showed 0.69 and 0.87 f1-scores, respectively (21% big difference in performance). For the captions, the group with the fewest (0<captions≤5) and that with the most (captions>5) showed 0.63 and 0.67 f1-scores, respectively (5% difference). The performance of the models with the captions was even lower than that of the ones with the visual hints. By and large, the models with the visual hints showed similar performances in any group (2% difference).

These results indicate that text information well represents user emotion in social media posts, yet the model performance with text information highly depends on its amount. Conversely, the performance dependency of the model with the visual hints on the amount of text information was low. Even the performance of the model learned from the posts with no text information was still high (0.71 f1-score). These results again confirmed the importance of visual hints to understand one's emotions expressed in social media posts. The results also ask for identifying additional characteristics of visual hints and applying newly derived features to emotion prediction. This effort can be one way to reflect recent trends in generation and sharing of social media posts (i.e., the increase use in images and the decrease use in text).

5 DISCUSSIONS

5.1 Color distribution

Our analysis showed that the distributions of the colors were not significantly different between the positive and negative emotions. Interestingly, these results are inconsistent with prior studies [1, 34] and can be explained by the nature of social media. Social media platforms are designed to support information sharing. A great number of posts come from individual users and contain information that reflects various aspects of a user's life (e.g., everyday activities, social interactions, personal interests), thought or even random images from the web [10]. Our literature review indicated that many prior studies used images of artistic photos or abstract paintings, which were mostly created by professionals. Mostly, those images are different from casual images of social media because there are much more freedom and dynamics in social media regarding information creation and sharing. Conventional notions may not be well applied to social media contexts.

5.2 Visual hint distribution

In the case of visual hints, the frequency of the topics and the visual hints of each topic appeared different between the positive and

negative emotions. Among the positive emotions, the most visual hints in happiness and excitement were related with `fashion`, and those in satisfaction were `food`. For the visual hints of the fashion topic, there were specific items, such as shoes, hats, jackets, and accessories. Such items could be the ones that the users purchased or received as a gift and from which they shared their positive emotion. Similarly, many visual hints on the `food` topic in satisfaction can be interpreted as sharing one's positive experience in dining and social interactions. Conversely, among the negative emotions, the topic of most common visual hint for sadness was `object`, that for fear was `color`, and that for anger was `state`. It is worth noting that those hints were somewhat abstract.

Based on these results, we could conclude that visual hints of images well represent a user's positive or negative emotions. In general, the positive emotions consist of explicit information (i.e., items in the images), whereas the negative emotions consist of implicit information (i.e., mood, state, color). These results are also consistent with our survey results (Figure 5): expressing one's emotion through the objects and the atmosphere in the images. Additionally, there were many visual hints about `person`, `indoor`, and `outdoor` in both emotions. These results seem reasonable, considering that one's social media image is a personal space for describing his/her daily lives. As recent review [45] highlighted a potential of contextual information of images, our study provides explicit insights of such information regarding emotions in online space.

5.3 Feature validation

Regarding the performance of emotion classification, the model with the visual hint features exhibited higher performance than the ones with the conventional color features. The FCN-based model using the visual hint features also showed higher performance than other models based on fine-tuning and transfer learning. These results indicate that the visual hints on which our study focused have greater correlation in expressing emotions than the raw features of images. Interestingly, even if we used all features from the images, the performance of the model did not increase compared to that of the model with the visual hints only. Given that the visual hints have the `color` topic, the general color information of the image may already be reflected in the visual hints.

Regarding the text features, the hashtags showed the best performance in the FCN-based model (0.74 f1-score). As highlighted in prior studies [2], we again identified a strong correlation between hashtags and user emotions. Regarding the methods of word representation, no difference was found in performance between Word2Vec and TF-IDF. From the results, we can infer that people tend to use hashtags without considering the order and use captions without considering the structure of the sentence.

In our dataset, hashtags have the largest portion in the text. However, the performance degraded as the number of hashtags and captions decreased. This result indicates that relying only on text information for emotion modeling may not work as expected if the amount of text information is insufficient.

In summary, the visual hint features showed their effectiveness and robustness in emotion classification. Because prior studies considered different emotion types with various datasets, it is challenging to directly compare the performance of our model with that of other studies. A key takeaway is that *images themselves play a significant role in implying emotions*. We demonstrated that the emotion classification model using only the visual hint features showed a comparable performance to the model using the text features (if text information is sufficient). As we previously emphasized, the image has become the primary information type in social media posts, and people's use and reliance on text is decreasing. Thus, such a result of comparable performance highlights *the importance and role of visual hints in understanding a user's emotion*. Our study results emphasize the necessity of analyzing emotions through image information and of eliciting salient features of the image, such as visual hints validated in our study, for better capturing users' emotions in the context of social media.

5.4 Limitations and future work

Our study results and analyses may not be generalized because we only considered Instagram as an image-based social media platform and the size of the dataset may not be large enough. However, our data collection was rigorously done by excluding the homogeneity of posts and filtering spams as much as possible. We believe that our direction for understanding the emotions of posts in social media is still valid. Yet, we plan to present more comprehensive results with more emotion-labeled data collected from additional image-based platforms. In addition, we focused on identifying various features from the image and the text that can be extracted from social media posts, which were found to be effective in understanding user emotions. However, our study needs further investigation in a methodological perspective (e.g., additional salient features from images, design of an optimal learning method or model structure), and there is room for the identification of additional visual hints, development of model algorithms, and improvement of the model.

6 CONCLUSION

In this paper, we strived to (1) reflect current trends in the creation of social media posts—an increase in image use and a decrease in text use—and (2) address limitations of prior research—the use of the insufficient image information—on emotion modeling. Because people often express their emotions in objects, we proposed the notion of *visual hints* as the contextual features of the image of social media posts. Our experimental results showed that (1) visual hints of each of the positive and negative emotions were different by topic and (2) the model with the visual hints exhibited a comparable performance to that with text features and a better performance than that with other raw image features. Finally, we discussed the results in connection with the characteristics of social media platforms. We demonstrated the validity of visual hints for user emotion analysis and modeling, in that the visual hints can compensate text information and improve the use of image features. Our study methods are expected to be used in various fields, such as online marketing recommendation, opinion mining, psychological healthcare, where the use of images and the importance of user emotions are high.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF, No. NRF-2020R1A2B5B03001960) and Institute of Information & Communications Technology Planning & Evaluation (IITP, No. 2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University)) grant funded by the Korea government (MSIT).

REFERENCES

[1] Barry J Babin, David M Hardesty, and Tracy A Suter. 2003. Color and Shopping Intentions: The Intervening Effect of Price Fairness and Perceived Affect. *Journal of Business Research* 56 (2003), 541–551.

[2] Matteo Baldoni, Cristina Baroglio, Viviana Patti, and Paolo Rena. 2012. From Tags to Emotions: Ontology-driven Sentiment Analysis in The Social Semantic Web. *Intelligenza Artificiale* 6 (2012), 41–54.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.

[4] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. 2013. Sentibank: Large-scale Ontology and Classifiers for Detecting Sentiment and Emotions in Visual Content. In *Proc. ACM Int'l Conf. on Multimedia*.

[5] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-Scale Visual Sentiment Ontology and Detectors Using Adjective Noun Pairs. In *Proc. ACM Int'l Conf. on Multimedia*.

[6] David B Bracewell. 2008. Semi-Automatic Creation of An Emotion Dictionary Using Wordnet and Its Evaluation. In *Proc. IEEE Int'l Conf. on Cybernetics and Intelligent Systems*.

[7] Lijuan Cai and Thomas Hofmann. 2003. Text categorization by boosting automatically extracted concepts. In *Proc. ACM SIGIR Conf. on Research and Development in Informaion Retrieval*.

[8] Ming Chen, Lu Zhang, and Jan P Allebach. 2015. Learning Deep Features for Image Emotion Classification. In *Proc. IEEE Int'l Conf. on Image Processing*.

[9] Mihaly Csikszentmihalyi and Eugene Halton. 1981. *The Meaning of Things: Domestic Symbols and The Self.* Cambridge University Press.

[10] Munmun De Choudhury, Scott Counts, and Mary Czerwinski. 2011. Identifying relevant social media content: leveraging information diversity and user cognition. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*. 161–170.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition & Emotion* 6 (1992), 169–200.

[13] Carroll E Izard. 1991. *The Psychology of Emotions.* Springer Science & Business Media.

[14] Carroll E Izard. 2007. Basic Emotions, Natural Kinds, Emotion Schemas, and A New Paradigm. *Perspectives on Psychological Science* 2 (2007), 260–280.

[15] Bo Jiang, Yun Ling, and Jiale Wang. 2010. Tag Recommendation Based on Social Comment Network. *International Journal of Digital Content Technology and its Applications* 4 (2010), 110–117.

[16] Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. 2015. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd ACM international conference on Multimedia*. 159–168.

[17] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR, 1188–1196.

[18] Xin Lu, Reginald B Adams, Jia Li, Michelle G Newman, and James Z Wang. 2017. An Investigation into Three Visual Characteristics of Complex Scenes that Evoke Human Emotion. In *Proc. IEEE Int'l Conf. on Affective Computing and Intelligent Interaction*.

[19] Xin Lu, Poonam Suryanarayan, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. 2012. On Shape and The Computability of Emotions. In *Proc. ACM Int'l Conf. on Multimedia*.

[20] Jana Machajdik and Allan Hanbury. 2010. Affective Image Classification Using Features Inspired by Psychology and Art Theory. In *Proc. ACM Int'l Conf. on Multimedia*.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[22] Genevieve Patterson and James Hays. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2751–2758.

[23] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proc. Conf. on Empirical Methods in Natural Language Processing*.

[25] Gerard Salton and Michael J McGill. 1986. *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc.

[26] Aaron Smith, Monica Anderson, and Tom Caiazza. 2018. Social Media Use in 2018.

[27] Qi Su, Xinying Xu, Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen, and Zhong Su. 2008. Hidden sentiment association in chinese web opinion mining. In *Proc. ACM Int'l Conf. on World Wide Web*.

[28] H. Tankovska. 2021. Global social networks ranked by number of users 2021.

[29] H. Tankovska. 2021. Social media platforms growth of MAU worldwide 2019-2021.

[30] Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. 40th Annual Meeting on Association for Computational Linguistics*.

[31] Peter D Turney and Michael L Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems* 21, 4 (2003), 315–346.

[32] Princeton University. 2010. About WordNet. *WordNet* (2010).

[33] Kate Vaisutis, Margot Brereton, Toni Robertson, Frank Vetere, Jeannette Durick, Bjorn Nansen, and Laurie Buys. 2014. Invisible Connections: Investigating Older People's Emotions and Social Relations Around Objects. In *Proc. ACM CHI Conf. on Human Factors in Computing Systems*.

[34] Patricia Valdez and Albert Mehrabian. 1994. Effects of Color on Emotions. *Journal of Experimental Psychology: General* 123 (1994), 394.

[35] Maria V Valueva, NN Nagornov, Pave A Lyakhov, Georgiy V Valuev, and Nikolay I Chervyakov. 2020. Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation* 177 (2020), 232–243.

[36] Pu Wang and Carlotta Domeniconi. 2008. Building semantic kernels for text classification using wikipedia. In *Proc. ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data mining*.

[37] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proc. AAAI Conf. on Artificial Intelligence*.

[38] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. 2006. Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models. *ACM Transactions on Asian Language Information Processing* 5 (2006), 165–183.

[39] Can Xu, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li. 2014. Visual Sentiment Prediction with Deep Convolutional Neural Networks. *arXiv preprint arXiv:1411.5731* (2014).

[40] Peng Yang, Qingshan Liu, and Dimitris N Metaxas. 2010. Exploring facial expressions with compositional features. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2638–2644.

[41] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark.. In *Proc. AAAI Conf. on Artificial Intelligence*.

[42] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. 2013. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*. 1–8.

[43] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring Principles-of-Art Features for Image Emotion Recognition. In *Proc. ACM Int'l Conf. on Multimedia*.

[44] Sicheng Zhao, Hongxun Yao, Yue Gao, Guiguang Ding, and Tat-Seng Chua. 2016. Predicting Personalized Image Emotion Perceptions in Social Networks. *IEEE Transactions on Affective Computing* (2016).

[45] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller, and Kurt Keutzer. 2021. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).