Article

Inferring Emotion Tags from Object Images Using Convolutional Neural Network

Anam Manzoor ¹, Waqar Ahmad ^{1,*}, Muhammad Ehatisham-ul-Haq ^{1,*} , Abdul Hannan ²,
Muhammad Asif Khan ¹ , M. Usman Ashraf ², Ahmed M. Alghamdi ³ and
Ahmed S. Alfakeeh ⁴

¹ Department of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan; anammanzoor3333@gmail.com (A.M.); masif.khan@uettaxila.edu.pk (M.A.K.)

² Department of Computer Science, University of Management and Technology, Lahore (Sialkot) 51040, Pakistan; abdul.hannan@skt.umt.edu.pk (A.H.); usman.ashraf@skt.umt.edu.pk (M.U.A.)

³ Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia; amalghamdi@uj.edu.sa

⁴ Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; asalfakeeh@kau.edu.sa

* Correspondence: waqar.ahmad@uettaxila.edu.pk (W.A.); ehatishamuet@gmail.com (M.E.-u.-H.)

Received: 5 July 2020; Accepted: 28 July 2020; Published: 1 August 2020

Abstract: Emotions are a fundamental part of human behavior and can be stimulated in numerous ways. In real-life, we come across different types of objects such as cake, crab, television, trees, etc., in our routine life, which may excite certain emotions. Likewise, object images that we see and share on different platforms are also capable of expressing or inducing human emotions. Inferring emotion tags from these object images has great significance as it can play a vital role in recommendation systems, image retrieval, human behavior analysis and, advertisement applications. The existing schemes for emotion tag perception are based on the visual features, like color and texture of an image, which are poorly affected by lightning conditions. The main objective of our proposed study is to address this problem by introducing a novel idea of inferring emotion tags from the images based on object-related features. In this aspect, we first created an emotion-tagged dataset from the publicly available object detection dataset (i.e., “Caltech-256”) using subject evaluation from 212 users. Next, we used a convolutional neural network-based model to automatically extract the high-level features from object images for recognizing nine (09) emotion categories, such as amusement, awe, anger, boredom, contentment, disgust, excitement, fear, and sadness. Experimental results on our emotion-tagged dataset endorse the success of our proposed idea in terms of accuracy, precision, recall, specificity, and F1-score. Overall, the proposed scheme achieved an accuracy rate of approximately 85% and 79% using top-level and bottom-level emotion tagging, respectively. We also performed a gender-based analysis for inferring emotion tags and observed that male and female subjects have discernment in emotions perception concerning different object categories.

Keywords: emotion recognition; convolutional neural network; deep learning; image emotion analysis; object categorization

1. Introduction

Emotions represent the mental and psychological state of a human being and are a crucial element of human daily living behavior [1,2]. They are a vital source to perceive, convey, and express a particular feeling and can stimulate the human mind 3000 times more than the rational thoughts [3]. With the rapid growth in digital media, emotion communication has increased through images [4–8],

texts [9], speech [10], and audio-video contents [11]. According to [12], the contribution in the field of emotion classification is mainly based on image, speech, video, and text stimuli to extract useful features for testing various kinds of recognition algorithms. Recent studies have also focused on recognizing emotions from visual stimuli that encompass different objects [13,14]. In real-life, the exposure of emotion occurrences can be either through digital media contents like visual stimuli or real-life objects [15]. However, it is tempting to ask what kind of information one can get from these visual stimuli to verify the image information interpretation. Likewise, it is crucial to know what type of emotions can be assigned to an image. There is no existing theory to model human emotions from images. However, according to the psychological and art theory [5], human emotions can be classified into specific types, such as dimensional and categorical. The dimensional theory represents the emotions as the coordinates of two or three-dimensional space, i.e., valence-arousal or valence-arousal-dominance [16]. On the other hand, categorical theory classifies the emotions into well-suited classes, e.g., positive and negative categories [6]. P. Ekman et al. [17] introduced five essential emotion labels based on facial expression, which include happiness, disgust, anger, fear, and sadness, without any discrete emotion classification. The same emotion categories have been utilized in various studies [18–20]. The authors in [21] presented eight emotion categories to study human emotion perception, where happiness emotion is split into four positive emotions, namely amusement, excitement, awe, and contentment, while negative sentiments are further categorized as anger, fear, disgust, and sad emotions [14,22]. In [3], nine emotion categories were introduced with the addition of boredom emotion for visual emotion analysis.

An image can be either interpreted in terms of explicit or implicit information. The explicit information is object recognition, whereas implicit information is known as emotion perception, which is subjective and more abstract than explicit information [23]. Moreover, there are two major types of emotions, including perceived emotion and induced emotions, as reported by P.S. Juslin [24]. The perceived emotion refers to what a recipient/viewer thinks that the sender wants to express. On the other hand, induced emotion refers to what a person perceives when subjected to some visual stimuli, such as an image or video. Thus, inferring induced emotions is somewhat difficult than perceived emotions [3]. Various researchers have devoted considerable efforts towards image emotion analysis using visual stimuli (i.e., images) incorporating machine learning [13,25–27] and deep learning [14,23,28] approaches. These studies aim to focus on modeling emotions based on several learning algorithms and to extract useful features for image emotion analysis. However, the problem of inferring emotion tags with hand-crafted visual features is very challenging as these features at times are not able to interpret a specific emotion from an image with high accuracy. It is because these features are based on human assumptions and observations and may lack some essential factors of modeling and image [14]. As a result, deep learning with a convolutional neural network (CNN) has shown greater success with automatically learned features in affective computing [29]. CNN-based models tend to produce a more accurate and robust result as compared to the standard machine learning algorithms in many computer vision applications. These applications include digit recognition, speech/pronunciation and audio recognition, semantic analysis, and scene recognition [30–35]. The research study in [28] demonstrated that deep learning-based CNN architecture is the most astonishing and off-the-shelf tool for feature extraction.

Generally, the information contained in the images includes diverse kinds of living object categories, such as flowers, birds, animals, and insects, etc. In contrast, the categories of the non-living objects included in the images contain household items, vehicles, buildings, sceneries, etc. According to [36], these objects divulge various types of emotions on human beings, and a person might have some emotional reactions when interacting with these objects. For instance, an image that contains some eatable items, like pizza, sushi spaghetti, etc., generally perceives emotions, such as excitement, amusement, or disgust. Likewise, a person may perceive an emotion such as contentment, excitement, or amusement while watching an image of a flower, tree or beautiful scenery. Similarly, a person can regard an emotion tag as either amusement by viewing a picture of a snowmobile or a heavy motorbike.

Hence, the exclusive information about a person's behavior can be acquired by interpreting their emotion perception towards different real-life object categories. In this aspect, user demographic information can also be taken into account for emotion recognition. The psychological and behavioral studies proved that the human perception of emotion varies concerning user demographics [8]. Only a few research studies have shown interest in classifying emotions based on user demographics that generally refer to user attributes such as age, gender, marital status, and profession. In [37,38], the authors reported a gender discrimination scheme for image-based emotion recognition. Dong et al. [39] analyzed the demographics that are widely used to characterize customer's behaviors. Likewise, Fischer et al. [38] reported that diverse genders and age groups have different emotion perceptions. Besides, men invoke powerful emotions (such as anger), and women have powerless emotions (such as sadness and fear). Likewise, the authors in [37] mentioned that females are more expressive in emotions as compared to males. Therefore, it can be realized that multiple emotion classification based on object-related features and demographic analysis is a challenging task.

For addressing the challenges discussed above, in this study, we explore the implications of inferring emotion tags from images based on object analysis and categorization. Human-object interaction (either in real-life or through object images) entails varying emotion perceptions for different users, as shown in Figure 1. Different objects may produce varying emotions on the viewers depending upon their judgment. Hence, it seems advantageous to infer a viewer's emotions based on visual object interaction through images. In this aspect, based on the majority voting principle, we can establish a standard way to annotate emotions associated with different object categories. Thus, our proposed scheme is envisaged on the hypothesis that we can accurately infer emotion tags from the images using object-related features to alleviate the existing challenges associated with visual features (such as image contrast, brightness, and color information). Thus, to validate the proposed idea, we used a publicly available Caltech-256 dataset that consists of a hierarchy of object categories, including 20 top-level and 256 bottom-level object categories. We pre-processed this dataset for assigning emotion tags to these object categories. To the best of our knowledge, no one else has reported such research work that infers the emotion tags from images based on object analysis.

Figure 1. A visual representation of emotion perception idea based on human-object interaction, where multiple sample users are interacting with real-life object images and perceiving diverse kind of emotions.

The main contributions of this paper are as follows:

- We proposed a novel idea for inferring emotion tags from images based on object classification. In this aspect, we used a public domain object detection dataset, i.e., Caltech-256, for implementation and validation of our proposed idea using different supervised machine learning classifiers.
- We produced an emotion-tagged dataset from Caltech-256, where we assigned overall nine emotion tags to different object categories based on subjective evaluation from 212 users. Thus, our proposed scheme can recognize nine (09) different emotions, including amusement, excitement, awe, anger, fear, boredom, sadness, disgust, and contentment.
- We employed a two-level (i.e., top-level and bottom-level) emotion tagging approach for our proposed scheme. We provided a detailed analysis of how human emotions are affected by a different level of information. Also, we investigated the effect of user demographics, i.e., gender, on inferring emotion tags.

The rest of the research study is organized as follows: Section 2 describes the methodology opted to accomplish the proposed idea. Section 3 entails the performance analysis, result and, discussion of inferring emotion tags using CNN-based features. Finally, Section 4 provides the conclusions for this research work and provide recommendations for future work.

2. Proposed Methodology

The proposed method for inferring emotion tags based on object categories consists of four steps, i.e., dataset acquisition, emotion annotation/tagging, feature extraction, and classification of emotion tags as shown in Figure 2. The detailed description related to each step is given below.

Figure 2. Block diagram of the proposed methodology. (a) Illustrates the subjective evaluation method applied to build an emotion-tagged dataset, and (b) shows the key steps taken to infer emotion tags, i.e., emotion classification.

2.1. Data Acquisition

The proposed work mainly emphasizes on the concept of inferring emotion tags from object categories. Thus, we required a dataset that contains real-life object categories that we encounter in our daily life. For this purpose, we utilized a state-of-the-art object-detection dataset named Caltech-256 [40]. Generally, this dataset comprises of 256 object categories and a total of 30,608 images. Moreover, each category has a maximum of 827 and a minimum of 80 images. Specifically, this dataset is chosen on account of two reasons: 1) the dataset contains most of the real-life living and non-living object categories, with which a person might have interactions in their routine life and, 2) the dataset includes those types of images that are excessively uploaded and shared on different social media platforms and may induce certain emotions on the viewers. Hence, this dataset is appropriate to validate the classification performance for inferring emotion tags based on object categories. Besides this, Caltech-256 dataset consists of a hierarchy of object categories, i.e., top-level (node-level) object categories that are twenty (20) in number, including animal, transport, electronics, nature, music and food categories, etc. Likewise, there 256 bottom-level (leaf-level) object categories, which are derived from twenty top-level object categories and entail no more sub-categories. As an example, the top-level animal object category further contains bottom-level categories like frog, snake, ostrich, and crab, etc. Figure 3, presents some of the example images of bottom-level object categories (represented by black color blocks), which are derived from the top-level object categories of Caltech-256 dataset (as a specimen represented by red color blocks).

Figure 3. Caltech-256 dataset: Example images of bottom -level object categories (highlighted in black boxes), which derived from top-level object categories (highlighted in red boxes).

2.2. Data Annotation and Tagging

The Caltech-256 dataset consists of object images that are generally desirable to use for object detection or recognition purposes [41]. According to the proposed scheme, the raw data acquired from the dataset need to be pre-processed to create an emotion-tagged dataset from the existing object categories in Caltech-256. Mainly, nine (09) emotion tags are targeted in this study, including amusement, anger, awe, boredom, contentment, disgust, excitement, fear, and sad. For assigning a particular emotion tag to the hierarchy of the objects, i.e., top-level and bottom-level object categories, a subjective evaluation method is devised. This method depends on viewers' opinion based on the majority voting criterion to build a strongly labeled emotion dataset from the heterogeneous object

categories of Caltech-256 dataset. For this purpose, we conducted an online survey from more than 250 subjects using Google forms, where we got complete responses from 212 subjects. The proportion of male and female candidates was nearly the same, i.e., 51.8% and 48.2%, respectively. The survey participants mostly included university-level students and teachers. A few other subjects, such as doctors and private job employees, were also a part of the survey. We asked these participants to view random images from each object category and label what kind of emotions they typically perceive when interacting with that specific category objects in real-life or on social media platforms. More precisely, we requested the viewers to assign a specific emotion tag to each object category based on their real-life interaction with such objects that are included in the Caltech-256 dataset. We firstly adopted a subjective evaluation method to label emotion tags for the bottom-level object categories. Then, we used those tags to assign emotion tags for twenty (20) top-level object categories using majority voting principle of sub-category labels. The detailed description of emotion tag assignment based on subjective evaluation (for both top and bottom-level object categories) is given in the next sections.

2.2.1. Emotion Tag Assignment to Bottom-Level Object Categories

To consider all possible emotion variations within the top-level object categories, we first performed emotion labeling at the bottom-level. In this regard, we labeled each of 256 leaf categories of the object in Caltech-256 dataset for emotion tagging. All users were asked to assign only one emotion tag to each category based on their perception. After collecting emotion responses concerning bottom-level object categories from all the users, we applied a maximum user response criterion for assigning a single emotion tag to each category. For example, some of the object categories like blimp, boom-box, comet, playing card, and dice are assigned with the amusement emotion tag as a result of the combined response from all users. Likewise, the categories of images such as picnic-table, raccoon, and Saturn are marked with contentment emotion tag. In contrast, object categories like knife, bear, and bat belong to the fear emotion class as labeled by most of the users. Subsequently, all the bottom-level object categories are marked with one out of nine emotion tags to build a strongly labeled object category dataset. Overall, our emotion-tagged dataset based on the bottom-level object categories includes a total of 6177 images for amusement, 277 for anger, 3292 for awe, 5792 for boredom, 4270 for contentment, 1569 for disgust, 6336 for excitement, 1674 for fear, and 356 images for sad emotion category, respectively.

2.2.2. Emotion Tag Assignment to Top-Level Object Categories

We employed a top-level emotion tag assignment to analyze the overall emotion response for twenty primary object categories included in Caltech-256. In this regard, we examined average responses concerning all emotion tags of bottom-level categories in relation to their parent categories, i.e., top-level categories. Finally, we marked the top-level categories with emotion tag based on the maximum occurrence of emotion from the respective bottom-level object categories. For example, a top-level object category such as animal encompasses three different bottom-level categories, including air, water, and land animals. Further, each of these animal categories consists of more animal categories related to their kinds, i.e., ostrich, crab, frog and, duck, etc. Accordingly, when we observed all the bottom-level object categories in animal class, most of the categories were marked with fear emotion tag. Thus, the top-level animal object category is assigned with the fear emotion tag. Similarly, the insect category is tagged with disgust emotion as most of the derived bottom-level object categories were labeled with the same emotion tag by the users. Likewise, food and extinct object categories are marked with excitement and sad emotion tags. In this way, the emotion tags are assigned to the all top-level object categories to produce an emotion-tagged dataset based on object categorization. The final dataset includes a total of 1456 images for amusement, 609 for anger, 991 for awe, 6814 for boredom, 2005 for contentment, 813 for disgust, 12,127 for excitement, 4698 for fear, and 190 for sad emotions. Table 1 presents the emotion tag assigned to 256 bottom-level object categories derived from the 20 top-level object categories.

Table 1. Subjective labeling of bottom-level object categories with their respective emotions tags.

Emotion	Object Categories Associated with Different Emotion Tags
Amusement	Baseball-bat, sunflower, Base-ball glove, steering wheel, Basket-ball hoop, blimp, swan, bonsai, cactus, cartman, sushi, chimp, cormorant, desk globe, duck, elephant, elk, galaxy, goldfish, hamburger, ostrich, hibiscus, homer Simpson, hummingbird, ibis, iris flower, palm tree, Jesus Christ, kangaroo, killer whale, lighthouse, people, pram, teddy bear, watch, waterfall
Anger	Ak47, brain, harpsichord, dolphin, traffic light
Awe	Butterfly, Conch, football helmet, goat, gorilla, awe, owl, unicorn
Boredom	Backpack, chopstick, coin, stained glass, computer keyboard, computer monitor, computer mouse, clutter, tennis shoes, yo-yo, spoon, tweezer, doorknob, PCI card, face easy, tambourine, drinking straw, eyeglasses, stirrups, palm pilot, Swiss army knife, fire extinguisher, socks, fire hydrant, VCR, flashlight, projector, floppy disk, necktie, xylophone, fried egg, paper clip, frying pan, gas pump, license plate, light bulb, telephone box, hammock, hourglass, theodolite, mandolin, rotary phone, microscope, photocopier, screwdriver, mandolin, Piz-dispenser, Saturn, screwdriver, Segway, self-propelled lawnmower, sextant, teapot, wheelbarrow, washing machine
Contentment	Bathtub, Contentment, Buda, calculator, car tire, dog, ewer, fern, Grapes, harmonica, ladder, lathe, refrigerator, tomato, yarmulke
Disgust	An American flag, bat, Disgust, cockroach, frog, horseshoe crab, hot dog, house fly, iguana, toad, mussels, octopus, skunk, snail, trilobite, wine bottle
Excitement	Billiards, binocular, birdbath, boom-box, bowling ball, bowling pin, bowling glow, cake, spaghetti, camel, cannon, canoe, CD, male box, Cereal box, chandelier, soda can, snowmobile, mushroom, chessboard, superman, coffee-mug, comet, covered wagon, cow-hat, diamond ring, mattress, dice, speed boat, megaphone, soccer ball, tower Pisa, treadmill, T-shirt, zebra, sneaker, dolphin, touring bike, top-hat, dump bell, Eiffel tower, electric guitar, fighter jet, fireworks, French horn, Frisbee, tricycle, giraffe, umbrella, tripod, golden gate bridge, paper shredder, tuning fork, rainbow, laptop, joystick, ice cream cone, golf ball, Kayak, goose, grand piano, guitar-pick, hot air balloon, headphone, hawkbill, telescope, roulette, school bus, wheel, saddle, tennis ball, skateboard, helicopter, horse, hot tub, iPod, picnic table, Ketch, porcupine, penguin, playing cards, skyscraper, smokestack, teepee, tennis court, tennis racket, toaster, watermelon, welding mask, windmill
Fear	Bear, bulldozer, spider, Centipede, fire truck, grasshopper, greyhound, snake, human skeleton, knife, leopards, lightning, llama, mantis, raccoon, revolver, rifle, scorpion, syringe, Triceratops
Sad	Coffin, tombstone

2.2.3. Gender-Based Emotion Tag Assignment

For gender-based emotion analysis, we separated the emotion tags reported by males and females participants and performed new annotation for emotion tags based on gender information. The primary purpose of gender-based emotion tagging is to investigate the benefits of employing user demographics for inferring emotion tags based on object categories. The subjective labeling consisted of 51.8% male and 48.2% female responses for all object categories. However, the responses were different for male and female candidates. For example, the bottom-level object category such as necktie is tagged with boredom sentiment based on the combine responses from all females. However, males marked the necktie object category with excitement emotion tags most of the times. Similarly, females marked spaghetti as excitement, while males tagged it with boredom. Hence, there seems to be an apparent advantage of using gender-based emotion analysis, which is the improvement in emotion classification performance. For assigning gender-based emotion tags at both top and bottom levels, we devised the same procedure that we opted for combined emotion responses from all the users.

2.3. Feature Extraction Using Off-the-Shelf CNN: Resnet-50

For accurately inferring emotion tags from different image categories based on object analysis, the selection of appropriate features is crucial. However, it is hard to know what features can produce good results, particularly in this case where heterogeneous images are present with a single emotion class. Each object category, tagged with a single emotion label, entails various types of objects with different appearances and shapes. Hence, it is not effective to utilize object-related features for emotion recognition as multiple object types may have diverse features within the same category. Similarly, visual features like color and texture of an image are generally extracted based on the domain knowledge, i.e., human observations and assumptions. Hence, they do not encounter all the visual details [42]. Consequently, a major gap arises between what humans perceive and what we get from these hand-crafted features. Therefore, it is required to extract deep high-level features from different object images for better realization and representation of emotion. For our purposed scheme, we preferred CNN-based automatically extracted features for inferring emotion tags from object images. In this regard, we utilized a CNN model that is pre-trained on the benchmark dataset, i.e., ImageNet [28,31], entailing 1000 object categories with 1.2 million training images. From this large collection of images, CNN can learn rich feature representations for a wide range of images, which can outperform the hand-rafterd features. One such type of pre-trained state-of-the-art CNN architecture is ResNet-50 that consists of 152 layers [43]. This residual network makes the training phase easier as compared to other CNN models due to its substantially deeper network layers [44]. The first layer of the ResNet-50 network utilizes filters for capturing blob and edge features. These primitive features are then processed by deeper network layers to combine the early features to form higher-level image features, as shown in Figure 4. These higher-level features are better suited for recognition tasks as they combine all the primitive features into a richer image representation [45]. Thus, selecting the rightmost deep layer, just before the classification layer, is an effective design choice as it incorporates more rich information is a good place to start. Therefore, in our proposed model, we used the last fully-connected layer, i.e., ‘fc1000’, of ResNet-50 to automatically learn features from the emotion-tagged data. We obtained 1×1000 feature vector per sample. The main reason for adopting ResNet-50 architecture is that it has a minimal error rate (i.e., 3.57% on the ImageNet test set) as compared to other models.

Figure 4. Block diagram of ResNet-50 architecture, where the number on the top of blocks shows the number of repeating units for each block and the number inside the block represents the layer output size and channels.

2.4. Feature Classification

To infer emotion tags, validate the performance of the automatically learned features via CNN, we selected three different classifiers, including Support vector machine (SVM), Random Forest (RF), and K-nearest neighbor (KNN). The SVM classifier is a supervised learning algorithm that uses the labeled samples to find the difference between classes [46]. Another one, KNN, is also a widely used classifier in the decision making systems as it has a very simple algorithm with stable performance. It uses the number of nearest training samples of the feature domain for the feature classification [47]. Finally, RF is also a powerful classifier, which is based on the combination of decision trees to predict a specific class. Therefore, these three classifiers are reported as foremost in producing high accuracies both for balance and imbalanced datasets due to their efficient prediction [48]. We selected a k-fold

cross-validation scheme with $k = 10$ to maximize the use of available data for training and testing the performance of selected classifiers. The entire dataset is split into ten equal folds, out of which nine data splits are used for training purpose, and the remaining fold is used for testing instances.

3. Results and Discussion

This section explains the series of analysis along with comprehensive performance and empirical results to evaluate our proposed scheme for inferring emotion tag based on object categories. Besides, it explores the impact of user demographics for inferring emotion tags. The detail for each related subject is described in the following sections.

3.1. Method of Analysis and Classifiers

The proposed scheme for inferring emotions is evaluated on Caltech-256 dataset by using three classifiers as a baseline, i.e., K-nearest neighbor (K-NN), Random Forest (RF), and, Support vector machine (SVM). These classifiers are chosen due to their efficient performance for visual sentiment analysis. The performance of these classifiers is validated based on the widely used performance metrics, such as accuracy, F1-measure, precision, recall, and specificity. The proposed scheme focuses on object categories consisting of multi-level classification at two different levels. These two levels are termed as 1) top-level emotion tags, i.e., the emotion tag responses collected from top-level object categories like animal, food and, natural, etc., and 2) bottom-level emotion tags, i.e., the emotion tag responses collected from sub-object categories like air, water, and land animals. The subsequent sections provide detailed experimental results and performance analysis for both levels.

3.1.1. Performance Analysis for Inferring Emotion Tags at Top-Level

For analysis of the top-level emotion inferring, the chosen classifiers are trained to infer nine emotion categories. As discussed earlier, the collected responses from subjective labeling divided into three types of analyses for the top-level object categories. The illustration of the above three categories are as follows:

- Combine analysis of inferring emotion tags: It is the analysis based on combined responses from both gender for Caltech-256 top-level object categories
- Analysis of inferring emotion tags for males: These are emotion responses that are collected from only males
- Analysis of inferring emotion tags for females: In this category, emotion responses are collected from only females.

After grouping these emotion responses into three categories, we measure the proposed scheme performance for each category. Table 2 provides the average performance analysis for inferring emotion tags using K-NN, RF, and SVM classifiers. These results conducted with the help of three different analyses, i.e., male, female, and combine emotion responses. It is observed from Table 2 that for all three types of analysis at top-level, the K-NN classifier outperforms RF and SVM classifiers. However, SVM performs well in the cases when the margin between different class boundaries, i.e., hyperplanes, is sufficiently large. However, it is not possible in the case of our proposed scheme with small inter-class variations. In addition, the SVM classifier does not perform well if the dataset contains noise, such as labeling noise. In that case, target classes seem to be overlapping. Moreover, RF is an ensemble classifier that builds a collection of random decision trees (with random data samples and subsets of features), each of which individually predicts the class label. The final decision is made based on the majority voting principle. RF is inherently less interpretable than an individual decision tree. Hence, in the case of noisy-labeled data and an imbalanced dataset, RF may produce some random results. For our proposed scheme, the best accuracy is achieved by the K-NN classifier, and the worst performance is achieved using RF classifier for top-level inferring emotion tag. It can also be observed from the Table 2, inferring emotion tags based on male responses, we achieved the highest accuracy

of 85.01%, which is 0.15% and 2.78% higher than the female and combined emotion tag responses. Similarly, female emotion tag responses for inferring emotion tags are better than combine emotion tag responses. Likewise, the RF and SVM classifiers showed similar performance for individual male, female, and combine emotion tag responses. Hence, gender information is effective for inferring emotion tags.

Table 2. Result analysis of inferring emotion tag performance for three different classifiers at top and bottom-level using K-NN.

	Classifier	Demographic	Specificity	Precision	Recall	F-Measure	Accuracy
Inferring Emotion Tags at Top-level	KNN	Male	0.506	0.851	0.850	0.850	85.01%
		Female	0.507	0.849	0.849	0.849	84.86%
		Combine	0.514	0.825	0.822	0.823	82.23%
	SVM	Male	0.509	0.841	0.841	0.841	84.11%
		Female	0.510	0.835	0.835	0.834	83.48%
		Combine	0.513	0.823	0.827	0.824	82.67%
	RF	Male	0.667	0.793	0.769	0.758	76.93%
		Female	0.530	0.787	0.766	0.753	76.55%
		Combine	0.538	0.758	0.738	0.724	73.78%
Inferring Emotion Tags at Bottom-level	KNN	Male	0.790	0.523	0.789	0.790	79%
		Female	0.782	0.525	0.782	0.782	78%
		Combine	0.747	0.535	0.746	0.747	75%
	SVM	Male	0.729	0.541	0.728	0.729	73%
		Female	0.729	0.541	0.728	0.729	73%
		Combine	0.749	0.535	0.749	0.749	75%
	RF	Male	0.674	0.559	0.721	0.674	65%
		Female	0.662	0.563	0.667	0.662	66%
		Combine	0.664	0.562	0.718	0.664	64%

Figure 5 shows the F1-measure values obtained for inferring emotion tag using KNN classifier. It can be observed from Figure 5 that the F1-measure value for the emotion classes such as boredom and awe, the males and females emotion responses outperform than the combined emotion tag results. However, in the case of other emotion classes, gender and combine emotion tag responses showed a quite similar value of F1-measure. It can be noted from Table 2 that for the top-level emotion tag responses of males, the highest value of F1-measure is achieved using the K-NN classifier, which is 0.01% and 0.92% higher than the F1-measure of SVM and RF classifiers respectively.

Figure 5. Best-case average F1-score for inferring emotion tags at top-level with nine different emotions using K-NN.

Moreover, Figure 6 provides the confusion matrix for the best case top-level inferring emotion tag analysis obtained by K-NN for male, female, and combined emotion tag responses. It can be observed from Figure 6, in the case of male-based emotion tag analysis, 2318 amusement emotion instances are truly classified out of 2868 instances. On the other hand, for emotion tag responses from the combined category, 1182 out of 1440 instances are truly classified for the amusement emotion. Similarly, 153 instances are truly classified from 187 instances when inferring sad emotion tags for male emotion responses. In the same way, for the female and combined emotion tags, 151 and 155 instances are truly classified out of 187 instances, for the sad emotion class, respectively. Inferring emotions using K-NN classifier indicates that it is easier to predict amusement, awe, boredom, excitement, disgust, fear, contentment, anger and sad emotion tags based on object categories. Moreover it is observed Figure 5 that inferring gender-based emotions using different object categories is more effective as compared to combined emotion inferring. Therefore, combined emotion inferring results are not as effective as those based on female and male emotion responses. However, the overall inferring emotion performance is practicable and sufficient.

e1	2318	13	34	336	42	4	302	39	2	e1	2323	5	44	255	23	0	191	25	2	e1	1182	3	8	74	28	0	126	18	1
e2	17	394	8	63	22	1	90	7	0	e2	38	400	11	85	6	1	55	6	0	e2	22	402	5	75	12	1	80	5	0
e3	26	5	713	55	16	5	106	48	6	e3	48	3	913	62	38	3	88	43	4	e3	16	2	250	19	28	5	611	45	4
e4	214	33	46	8791	147	2	523	39	1	e4	235	48	106	6379	62	5	659	43	3	e4	74	41	51	5640	105	5	790	37	2
e5	21	4	6	88	1757	17	43	70	0	e5	57	7	33	69	528	0	56	7	0	e5	41	8	19	92	1584	20	145	75	0
e6	7	2	6	8	35	695	9	41	1	e6	8	3	4	16	1	694	31	46	1	e6	9	2	7	11	28	696	8	42	1
e7	312	50	79	580	84	8	5466	62	4	e7	299	37	117	637	38	23	8752	143	2	e7	180	38	699	747	150	6	10088	94	3
e8	35	5	34	92	144	56	92	4173	19	e8	40	5	29	80	8	61	230	4180	17	e8	37	5	33	52	128	56	144	4175	20
e9	2	0	1	3	2	3	3	20	153	e9	1	0	2	3	0	3	6	21	151	e9	1	0	2	4	1	3	3	18	155
	e1	e2	e3	e4	e5	e6	e7	e8	e9		e1	e2	e3	e4	e5	e6	e7	e8	e9		e1	e2	e3	e4	e5	e6	e7	e8	e9
(a)										(b)										(c)									

Figure 6. Best-case confusion matrices obtained for inferring emotion tags based on top-level object categories using K-NN classifier with (a) male responses, (b) female responses, (c) combined responses. The columns represent the predicted values, while the rows represent the ground truths. Here, e1, e2, e3, ..., e9 represent Amusement, Anger, Awe, Boredom, Contentment, Disgust, Excitement, Fear, and Sad emotion, respectively.

3.1.2. Performance Analysis for Inferring Emotion Tags at Bottom-Level

This section provides the performance analyses of male, female, and combined emotion tag responses for bottom-level inferring emotions, as shown in Table 2. It is observed that the K-NN classifier is giving better results as compared to SVM and RF classifiers for three emotion tag responses. Likewise, the SVM classifier also gives better results for individual emotion tag male, female, and combined responses. On the other hand, the worst accuracy for three different scenarios is achieved with RF classifier, which is only 65% for inferring emotion tags based on male responses. It is investigated that gender-based emotion tag response performs better than the combined based emotion tag responses for K-NN, SVM, and RF classifier. This indicates that splitting the combined emotion tag responses into gender-wise emotion responses leads to better performance for inferring emotions. The best accuracy of 79% and 78% is achieved from male and female emotion tag responses, respectively. The SVM and RF classifiers depict the same performance for male, female and combine emotion tags based responses. For emotion tags at bottom-level, Figure 7 compares the best case F1-measure values of each of the emotion class achieved using K-NN. The F1-measure achieved for bottom-level emotion inferring with respect to male subjects is 0.789 using K-NN classifier, which is 0.042% and 0.147% higher than the SVM and RF classifiers, respectively. As shown in Figure 6, the F1-measure achieved for the boredom, contentment, disgust and sad emotion class for male-based emotion responses outperforms the female and combined emotion tag responses. Similarly, for the amusement, awe, and fear emotion class based on female responses, the value of F1-measure is higher than the male and combined emotion tag responses. On the other hand, the value of F1-measure achieved for the anger, and sad emotion class for the female-based emotion tag responses are low as compared to male and combined based emotion tag responses. Generally, females are fewer respondents to object categories includes in Caltech-256 object category dataset for the anger and sad emotion class as compared to males.

Figure 7. Best-case average F1-score for inferring emotion tags at bottom-level with nine different emotions using K-NN.

It can be observed from Figure 8, that when using K-NN classifier in the case of male-based emotion tag inference at the bottom-level, the amusement emotion tag is truly classified 1835 times out of total 2567 instances. Likewise, in the case of female emotion tag responses for the amusement emotion category, 4782 instances are truly classified out of 6255 instances. On the other hand, in the case of combined emotion tag responses for the amusement category, only 2638 instances are truly classified out of 3562 in total. Also, for the sad emotion class, 468 instances are truly classified out

of 571 instances of male responses. For female and combined responses of the sad emotion class, 97 instances out of 491 and 162 out of 272 instances are truly classified respectively. The effective results for inferring emotion tags using the K-NN classifier indicate that it is easier to predict amusement, awe, boredom, excitement, disgust, fear, and contentment sentiments except for anger and sad tag based on object categories. In particular, the sad and anger sentiments vary from object to object both for males and females. Since the objects in the bottom-level of the Caltech-256 dataset has fewer categories, which perceive anger and sad emotions to both female and male while labeling the object categories with emotion tags. Specifically, for both male and female, emotion perception varies for the same object categories included in the Caltech-256 dataset. Thus, the results for combined emotion inferring are not as effective as compared to the results obtained from male and female based emotion tag results particularly. However, the overall performance for emotion inferring is practicable and sufficient. Furthermore, it can be noted that the inferring emotion tag accuracy achieved from the top-level analysis is higher than the bottom-level emotion inferring. The reason for this is that when we go into the detail of an object category, we come up with more emotional responses as emotion response varies for sub-object categories. This indicates that inferring emotion tag performance varies when we perceive emotions from top to bottom level object categories.

e1	1835	10	19	323	14	7	315	29	15
e2	6	468	0	57	2	1	15	14	8
e3	29	3	412	71	3	5	53	11	3
e4	238	56	40	9001	106	90	852	82	55
e5	14	1	4	94	649	12	30	2	0
e6	13	0	9	170	16	1052	59	16	4
e7		23	82	1094	50	55	5498	75	31
e8	34	26	5	166	3	32	78	796	6
e9	13	11	9	114	1	6	49	6	355
	↖	↖	↖	↖	↖	↖	↖	↖	↖
(a)									
e1	2638	13	63	227	79	42	432	52	16
e2	13	157	10	40	3	0	59	96	0
e3	73	3	837	46	10	21	120	42	5
e4	175	26	44	4847	122	48	710	101	16
e5	42	4	12	149	1088	10	169	39	2
e6	69	1	32	94	33	1238	131	62	0
e7	398	48	97	811	188	101	9161	190	20
e8	58	121	29	141	34	47	208	1972	3
e9	15	1	3	30	12	2	46	1	162
	↖	↖	↖	↖	↖	↖	↖	↖	↖
(b)									
e1	4782	38	203	332	308	104	385	49	54
e2	30	82	35	13	27	1	13	3	69
e3	219	33	2467	133	119	27	188	65	7
e4	216	14	116	4375	210	94	481	34	93
e5	230	6	86	210	2996	77	242	48	6
e6	131	1	50	114	67	1033	109	53	2
e7	367	11	140	492	255	109	4710	67	9
e8	67	1	47	72	97	59	74	1218	21
e9	48	74	8	143	35	2	10	22	97
	↖	↖	↖	↖	↖	↖	↖	↖	↖
(c)									

Figure 8. Best-case confusion matrices obtained for inferring emotion tags based on bottom-level object categories using K-NN classifier with (a) male responses, (b) female responses, (c) combined responses. The columns represent the predicted values, while the rows represent the ground truths. Here, e1, e2, e3, . . . , e9 represent Amusement, Anger, Awe, Boredom, Contentment, Disgust, Excitement, Fear, and Sad emotion, respectively.

3.2. Comparison with the State-of-the-Art Study

This section provides a comparative analysis of the proposed scheme for inferring emotion tags with state-of-the-art techniques. A comparison of the proposed method with the well-known existing approaches is presented in Table 3. It can be noted from Table 3 that all the previous studies are based on the self-developed datasets, such as images obtained from Flickr, to accomplish a desirable inferring emotion tag performance. Moreover, these studies are based on visual features, such as context and content-wise analyses of visual stimuli. Consequently, these schemes perform poorly in real-world scenarios. Moreover, it is also hard to opt for the human behavioral analysis based on such schemes as most of the time, the visual information like color and texture of an image is not enough to perceive a specific emotion tag. However, the proposed scheme considers an image object category as a visual feature to perceive an emotion tag. Thus, it is capable of efficiently infer nine emotion classes based on three types of analysis, i.e., combined, female, and male emotion responses. The proposed scheme effectiveness can be demonstrated by comparing the output results for emotion inferring with the existing state-of-the-arts. It can be observed from Table 3 that the accuracy rates achieved for the proposed scheme outperform the reported results of the existing emotion recognition studies. The apparent advantages of our proposed scheme offer effective decision making over the existing schemes, which are crucial to opt for real-time applications.

Table 3. Comparison of the proposed method with some well-known deep learning-based inferring emotion tag schemes.

Study	Technique	Dataset	No. of Emotion Categories	Performance Parameters	Results
[12]	Noisy fine-tuned CNN, ImageNet CNN, fine-tuned CNN	Flickr dataset (23,000 images)	8	Accuracy	58% with fine tunes, 32% with ImageNet and 46% with Noisy fine-tuned
[23]	CNN, PCNN	Flicker data set (half a million images)	2	Precision, Recall, Accuracy, F1 measure	Accuracy is 72% with CNN and 78% with PCCN
[14]	Off the shelf CNN, CNN with multi-scale pooling	Art photo (806 images) + Flickr dataset (11,575 images)	8	Average true positive Rate	Average true positive rate is 80% with multi-scale CNN
Proposed	Off the shelf CNN with ResNet-50	Caltech-256 (30,608 images)	9	Accuracy, F1-measure, Precision, recall	Bottom level Accuracy is 79% precision is 0.789 The recall is 0.790, F1-score is 0.789.
	Off the shelf CNN with ResNet-50	Caltech-256 (30,608 images)	9	Accuracy, F1-measure, Precision, recall	Top-level Accuracy is 85%, precision is 0.851, The recall is 0.850, F1-score is 0.850.

Despite efficient recognition performance as compared to the existing studies, our proposed system entails certain limitations. For example, it only takes into consideration the gender information for emotion perception and analysis. However, emotions do vary for different human beings with respect to their regions, traditions, and other demographics. To be more precise, as we presented in Section 2.2.3, for example, the object “dog” is category classified under “contentment” emotion tag, which is highly diverse between individuals belonging to different region, age, and profession. We just use the majority voting in this case for assigning the emotion tags. Thus, the classification of emotion tags from images based on the proposed scheme suffers from the problem of subjectivity. Therefore, for accurately modeling human emotions based on object interactions, it is required to incorporate the detailed user demographics in the proposed model as well to generalize a large diverse population.

4. Conclusions

In this research work, we presented an efficient scheme for inferring emotion tag from object images. We pre-processed the Caltech-256 dataset and applied subjective evaluation from emotion labeling, in which we adopted the CNN model (ResNet-50) to extract deep features for emotion classification. A novel method is devised to infer emotion tags from visual stimuli based on image object categories rather than color and texture features. We trained and inferred nine emotion classes, including amusement, anger, awe, boredom, contentment, disgust, excitement, fear and, sadness, from object images. Moreover, we used gender-based analysis to see how emotion perception varies from males to females. Overall, the proposed method achieved an accuracy of approximately 85% for top-level and 79% for bottom-level emotion tag recognition, which showed the effectiveness of the proposed idea. In future, the proposed scheme can be expanded to take into consideration more user demographic information to better model and infer the human emotions. The proposed study and the output results can be utilized in the future studies for a wide-ranging application, such as recommender systems and emotion-based content generation and retrieval. Furthermore, considering the adaptability of the proposed mechanism, it can be used to depict or infer the mental state of a person, which in turn can be quite useful in medical fields for patient’s treatment. Moreover, the proposed scheme can be used to model human behavior based on human-object interactions and emotion

recognition. With the increasing use of visual contents on social media platforms, the proposed scheme can help in analyzing the behavior of different communities towards some particular problem.

Author Contributions: For research articles Conceptualization, A.M., M.E.-u.-H., A.H.; methodology, A.M., M.E.-u.-H.; software, A.M., A.H.; validation, W.A., M.E.-u.-H.; investigation, M.U.A., M.E.-u.-H., M.A.K.; resources, A.M.A.; writing—original draft preparation, A.M., M.E.-u.-H., A.H., W.A.; writing—review and editing, M.E.-u.-H., A.H., M.A.K.; visualization M.E.-u.-H., W.A.; supervision, W.A., A.K., M.U.A.; project administration, W.A., M.U.A.; funding acquisition, A.M.A., A.S.A.; All authors have read and agreed to the submitted version of the manuscript.

Funding: This research is funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. (RG-3-611-38).

Acknowledgments: This research is funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. (RG-3-611-38). The authors, therefore, acknowledge with thanks to DSR technical and financial support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Picard, R.W.; Rosalind, W. Toward agents that recognize emotion. *VIVEK-BOMBAY* **2000**, *13*, 3–13.
2. Izard, C.E. *The Psychology of Emotions*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1991.
3. Wu, B.; Jia, J.; Yang, Y.; Zhao, P.; Tang, J.; Tian, Q. Inferring emotional tags from social images with user demographics. *IEEE Trans. Multimed.* **2017**, *19*, 1670–1684. [[CrossRef](#)]
4. Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In Proceedings of the 21st ACM international conference on Multimedia, Barcelona, Spain, 21–25 October 2013.
5. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Trans. Multimed.* **2010**, *12*, 682–691. [[CrossRef](#)]
6. Siersdorfer, S.; Minack, E.; Deng, F.; Hare, J. Analyzing and predicting sentiment of images on the social web. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 715–718.
7. Li, B. Scaring or pleasing: Exploit emotional impact of an image. In Proceedings of the 20th ACM international conference on Multimedia, Nara, Japan, 29 October–2 November 2012; pp. 1365–1366.
8. Cai, W.; Jia, J.; Han, W. Inferring emotions from image social networks using group-based factor graph model Department of Computer Science and Technology. In Proceedings of the IEEE International Conference on Multimedia and Expo 2018, San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
9. Calix, R.A.; Mallepudi, S.A.; Chen, B.; Knapp, G.M. Emotion recognition in text for 3-D facial expression rendering. *IEEE Trans. Multimed.* **2010**, *12*, 544–551. [[CrossRef](#)]
10. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.M.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th International Conference on Multimodal Interfaces, ICMI 2004, State College, PA, USA, 13–15 October 2004; pp. 205–211.
11. Lin, J.-C.; Wu, C.-H.; Wei, W.-L. Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. *IEEE Trans. Multimed.* **2011**, *14*, 142–156. [[CrossRef](#)]
12. You, Q.; Luo, J.; Jin, H.; Yang, J. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *arXiv* **2016**, arXiv:1605.02677.
13. Machajdik, J.; Hanbury, A. Affective image classification using features inspired by psychology and art theory. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 83–92.
14. Chen, M.; Zhang, L.; Allebach, J.P. Learning deep features for image emotion classification. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 4491–4495.
15. Stephen, A.T. The role of digital and social media marketing in consumer behavior. *Curr. Opin. Psychol.* **2016**, *10*, 17–21. [[CrossRef](#)]

16. Plutchik, R. A general psychocvolutionary theory of emotion. *Emot. Theory Res. Exp.* **1980**, *1*, 3–31.
17. Ekman, P. An argument for basic emotions. *Cognit. Emot.* **1992**, *6*, 169–200. [[CrossRef](#)]
18. Wang, S.; Wang, J.; Wang, Z.; Ji, Q. Multiple emotion tagging for multimedia data by exploiting high-order dependencies among emotions. *IEEE Trans. Multimed.* **2015**, *17*, 2185–2197. [[CrossRef](#)]
19. Wang, X.; Jia, J.; Tang, J.; Wu, B.; Cai, L.; Xie, L. Modeling emotion influence in image social networks. *IEEE Trans. Affect. Comput.* **2007**, *6*, 1–13. [[CrossRef](#)]
20. Vanus, J.; Belesova, J.; Martinek, R.; Nedoma, J.; Fajkus, M.; Bilik, P.; Zidek, J. Monitoring of the daily living activities in smart home care. *Hum. Cent. Comput. Inf. Sci.* **2017**, *7*. [[CrossRef](#)]
21. Mikels, J.A.; Fredrickson, B.L.; Larkin, G.R.; Lindberg, C.M.; Maglio, S.J.; Reuter-Lorenz, P.A. Emotional category data on images from the international affective picture system. *Behav. Res. Methods* **2005**, *37*, 626–630. [[CrossRef](#)] [[PubMed](#)]
22. Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.; Sun, X. Exploring principles-of-art features for image emotion recognition. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014.
23. You, Q.; Luo, J.; Jin, H.; Yang, J. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial intelligence, Austin, TX, USA, 25–30 January 2015.
24. Juslin, P.N.; Laukka, P. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *J. New Music Res.* **2004**, *33*, 217–238. [[CrossRef](#)]
25. Zhao, S. Affective computing of image emotion perceptions. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; p. 703.
26. Wu, B.; Jia, J.; Yang, Y.; Zhao, P.; Tang, J. Understanding the emotions behind social images: Inferring with user demographics. In Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, Italy, 29 June–3 July 2015; pp. 1–6.
27. Lu, X.; Suryanarayan, P.; Adams, R.B.; Li, J.; Newman, M.G.; Wang, J.Z. On shape and the computability of emotions categories and subject descriptors. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012.
28. Sharif, A.; Hossein, R.; Josephine, A.; Stefan, S.; Royal, K.T.H. CNN features off-the-shelf: An astounding baseline for recognition. *arXiv* **2014**, arXiv:1403.6382. Available online: <https://arxiv.org/abs/1403.6382> (accessed on 30 July 2020).
29. Affective Computing. Available online: <http://en.wikipedia.org/wiki/Affectivecomputing> (accessed on 30 July 2020).
30. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*; IEEE: Montreal, QC, Canada, 1998; Volume 86, pp. 2278–2324.
31. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In Proceedings of the NIPS 2012: Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012.
32. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
33. Sergey, K.; Matthew, T.; Helen, H.; Aseem, A.; Trevor, D.; Aaron, H.; Holger, W. Recognizing image style. *arXiv* **2013**, arXiv:1311.3715. Available online: <https://arxiv.org/abs/1311.3715> (accessed on 30 July 2020).
34. Akhtar, S.; Hussain, F.; Raja, F.R.; Ehatisham-Ul-haq, M.; Baloch, N.K.; Ishmanov, F.; Bin Zikria, Y. Improving mispronunciation detection of arabic words for non-native learners using deep convolutional neural network features. *Electronics* **2020**, *9*, 963. [[CrossRef](#)]
35. Caterina, S.M.; Gainotti, G. Interaction between vision and language in category-specific semantic impairment. *Cogn. Neuropsychol.* **1988**, *5*, 677–709.
36. Demir, E.; Pieter, M.A.D.; Paul, P.M.H. Appraisal patterns of emotions in human-product interaction. *Int. J. Des.* **2009**, *3*, 2.
37. Fischer, A.H.; Mosquera, P.M.R.; van Vianen, A.E.M. Gender and culture differences in emotion. *Emotion* **2016**, *4*, 87–94. [[CrossRef](#)]
38. Simon, R.W. Gender and Emotion in the United States: Do men and women differ in self—Reports of feelings and expressive behavior? *Am. J. Sociol.* **2004**, *109*, 1137–1176. [[CrossRef](#)]

39. Dong, Y.; Yang, Y.; Tang, J.; Chawla, N.V. Inferring user demographics and social strategies in mobile social networks. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014.
40. Caltech256. Available online: <http://academictorrents.com/details/7de9936b060525b6fa7f5d8aabd316637d677622> (accessed on 30 July 2020).
41. Griffin, G.; Alex, H.; Perona, P. *Caltech-256 Object Category Dataset*; California Institute of Technology: Pasadena, CA, USA, 2007.
42. Luo, J.; Andreas, E.S.; Amit, S. A bayesian network-based framework for semantic image understanding. *Pattern Recognit.* **2005**, *38*, 919–934. [CrossRef]
43. CNN Networks. Available online: <https://medium.com/@sidereal/cnns-architectures-lexnet-alexnet-vgg-googlenet-resnet-and-more-666091488df5> (accessed on 30 July 2020).
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Las Vegas, NV, USA, 27–30 June 2016.
45. Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv* **2013**, arXiv:1310.1531. Available online: <https://arxiv.org/abs/1310.1531> (accessed on 30 July 2020).
46. Demirhan, A. Performance of machine learning methods in determining the autism spectrum disorder cases. *Mugla J. Sci. Technol.* **2018**, *4*, 79–84. [CrossRef]
47. Demirhan, A. Neuroimage-based clinical prediction using machine learning tools. *Int. J. Imaging Syst. Technol.* **2017**, *27*, 89–97. [CrossRef]
48. Noi, P.T.; Kappas, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors* **2018**, *18*, 18.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).