AffectI: A Game for Diverse, Reliable, and Efficient Affective Image Annotation

Xingkun Zuo
Hangzhou Dianzi University,
Hangzhou, China
University of Yamanashi, Kofu, Japan
luckykunzuo@gmail.com

Jiyi Li
University of Yamanashi, Kofu, Japan
jyli@yamanashi.ac.jp

Qili Zhou
Hangzhou Dianzi University,
Hangzhou, China
zql@hdu.edu.cn

Jianjun Li
Hangzhou Dianzi University,
Hangzhou, China
lijjcan@gmail.com

Xiaoyang Mao*
University of Yamanashi, Kofu, Japan
Hangzhou Dianzi University,
Hangzhou, China
mao@yamanashi.ac.jp

ABSTRACT

An important application of affective image annotation is affective image content analysis, which aims to automatically understand the emotion being brought to viewers by image contents. The so-called subjective perception issue, i.e., different viewers may have different emotional responses to the same image, makes it difficult to link image features with the expected perceived emotion. Due to the ability to learn features, recent deep learning technologies have opened a new window on affective image content analysis, which has led to a growing demand for affective image annotation technologies to build large reliable training datasets. This paper proposes a novel affective image annotation technique, AffectI, for efficiently collecting diverse and reliable emotional labels with the estimate emotion distribution for images based on the concept of Game With a Purpose (GWAP). AffectI features three novel mechanisms: a selection mechanism for ensuring all emotion words being fairly evaluated for collecting diverse and reliable labels; an estimation mechanism for estimating the emotion distribution by aggregating partial pairwise comparisons of the emotion words for collecting the labels effectively and efficiently; an incentive mechanism shows the comparison between current player and her opponents as well as all past players to promote the interest of players and also contributes the reliability and diversity. Our experimental results demonstrate that AffectI is superior to existing methods in terms of being able to collect more diverse and reliable labels. The advantage of using GWAP for reducing the frustration of evaluators was also confirmed through subjective evaluation.

CCS CONCEPTS

• **Information systems → Multimedia information systems**; • **Human-centered computing → Collaborative and social computing**.

KEYWORDS

affective image annotation, dataset creation, game with a purpose

ACM Reference Format:
Xingkun Zuo, Jiyi Li, Qili Zhou, Jianjun Li, and Xiaoyang Mao. 2020. AffectI: A Game for Diverse, Reliable, and Efficient Affective Image Annotation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413744

1 INTRODUCTION

Affective image annotation has a wide range of applications, such as image retrieval and affective image content analysis. For image retrieval, there are many cases in which people may want to retrieve an image that represents or can provoke a particular emotion. On the other hand, affective image content analysis is an emerging research topic aimed at automatically recognizing or understanding the emotion evoked in the viewers by the visual contents [29]. A major challenge of affective image content analysis is how to deal with the so-called subjective perception issue, i.e., different viewers may have completely different emotional responses to the same image, depending on their gender, personality, or cultural and social background. Such subjectivity makes it difficult to link image features with the emotion the viewer is expected to experience by perceiving the signal [16, 25, 30]. With its ability to learn features in an end-to-end fashion, deep learning technologies are gaining significant attention for affective image content analysis, which has led to a significant demand for affective image annotation technologies required for building large reliable training datasets.

To tackle the subjective perception issue, existing learning-based affective image content analysis methods have two main ways of establishing image-to-label mapping, i.e., single-label learning that assigns a single emotional label to each image, which can be used for predicting personalized or dominant (average) emotion perception, and multi-label learning [20], which associates multiple labels to

[*] Corresponding Author

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413744

each image. Although multi-label learning can solve many problems of ambiguity in labeling, the importance or degree of different labels is actually unequal. In other words, emotion perception is also relative. Emotion distribution learning, which aims to learn the degree to which each emotion is invoked by the image has been developed [8]. To meet the requirements of those state-of-the-art affective image content analysis technologies, a new affective image annotation technology should be able to associate each image with multiple labels as well as with the emotion distribution faithfully representing the degree of emotions of a large variety of viewers.

Existing large-scale affective image datasets are constructed mainly using images from social networks by combining natural language processing and manual labeling [2, 26, 27, 30]. For example, the FlickrCC dataset [2] is constructed by retrieving the Flickr creative common (CC) images with 3,000 adjective-noun pairs (ANPs). The MVSO dataset [11] is a multilingual version of the FlickerCC dataset. The FI dataset [27] was constructed by first searching Flicker and Instagram with emotion keywords and then having the weakly labeled images further labeled by 225 Amazon Mechanical Turk (AMT) workers. Although making use of the text information annotated to the images on the social network is efficient for collecting large-scale datasets, the data collected in this way are usually very noisy, as the emotional words included in the text descriptions may depend on some context not really related to the emotion being conveyed by the images. The collected labels are biased to the high frequency emotional labels and thus lack of diversity and quality.

In this paper, we propose a novel affective image annotation technique AffectI, for collecting emotional labels for images based on the concept of Game With a Purpose (GWAP). GWAP was first proposed by Ahn and Dabbis, who developed ESP [21] game for collecting labels for images. There are also several other existing works similar to ESP game, such as KKB [10], Phetch [22], Peekaboom [23], and all tend to collect the surface semantic descriptions of images. Karido [19] and Artigo [4] mentioned collecting the deep semantics of images, but all these games can only collect descriptions about the object in the image, such as a bag, a car, a person, or some scenery represented by the image. To the best of our knowledge, no game is currently available for affective image annotation with multi labels as well as the emotion degree of the labels.

The proposed system in this paper mainly has three components, i.e., *selection mechanism* (Figure 1.(a)) which ensures that all emotion words being fairly evaluated and assists the players to select and judge the emotion degree of words in an efficient way; *estimation mechanism* (Figure 1.(b)) which estimates the emotion degree of all labels from partial rank list proved by multiple players; *incentive mechanism* (Figure 1.(c) and (d)) which shows the comparison between current player and her opponent or all past players.

The system concentrates on *efficiently* collecting *diverse* and *reliable* emotional labels for images. First, it can collect *diverse* affective labels with *diverse emotion degrees* rather than only using the normal emotion words with high frequency like that in the existing work [11]. Second, the diverse emotional labels that we collect have *high reliability*; the *emotion distribution* of all emotion words in an emotion taxonomy (Plutchik's Wheel of Emotions [17]) are provided; the system is thus *reliable* for affective image content analysis. Third, the players only need to partially judge the emotion degrees of small subsets in all emotion words which can decrease the workload of players; AffectI enables players to label the images in a way that is intuitive and with pleasure. The system is thus *efficient*. The main contributions are as follows.

- We propose a novel affective image annotation system, namely AffectI, which can efficiently collect reliable and diverse emotional labels with emotion degrees of these labels for images.
- The selection mechanism makes all candidate emotion words shown with fair opportunities and contributes the diversity and quality of the labels; the estimation mechanism estimates the emotion distribution of image from partial judgements and thus contributes the reliability and efficiency.
- The incentive mechanism can encourage the player to provide labels with high quality. We also empirically find that the past player based incentive also promote the players to provide diverse labels.
- We construct a novel image dataset with emotion distribution resulting from our experiment which can be used for affective image content analysis.

2 RELATED WORK

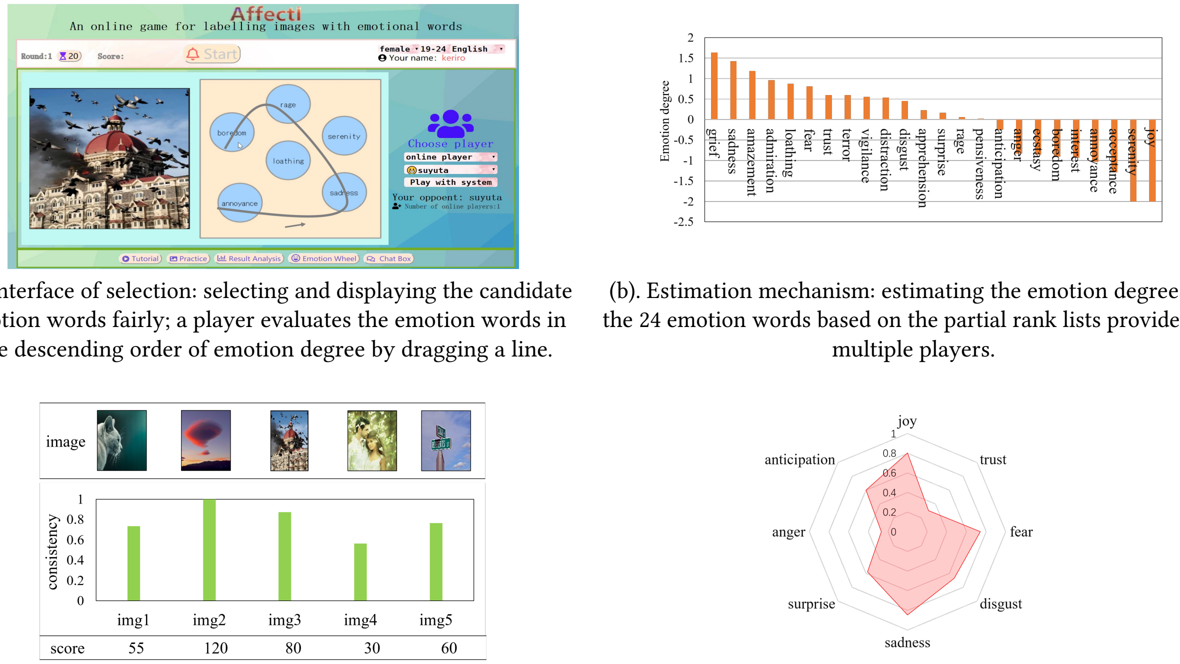
2.1 Emotion models

There are mainly two approaches to model emotions. One is coordinate based and the other is category based. Coordinate based approach uses 3D or 2D Cartesian space to represent emotions [9, 18]. Valence-arousal-dominance [18] together with its 2D version[9] is the most widely used coordinate based model, where valence represents the pleasantness, arousal the intensity of emotion and dominance the degree of control. Coordinate based model provides a continuous representation to emotions and hence is mostly used for regression tasks. Category based approach classifies emotion into a few basic categories, such as joy, anger, fear, etc. Well known models include Ekman's six basic emotions [7], Mikels's eight emotions [14] and Plutchik's Wheel of Emotions [17] which has 8 primary categories, each of which has three different degrees from strong to weak. Since the existing emotion label and distribution learning all use category based models, we adopt Plutchik's Wheel of Emotions, the most comprehensive category based model in AffectI.

Because an image can stimulate different emotions of varying degrees, it is important to identify the degree of each emotion. This is equivalent to define an emotion distribution for an image. Predicting such emotion distribution rather than a single dominant emotion of image has become to be the main stream of current affective image contents analysis studies [8, 25, 28]. Probability models are typically used to calculate emotion distribution. Two main forms of emotion distribution are used in existing studies: discrete distribution in category based analysis and continuous distribution in coordinate based analysis. Since our game is based on Plutchik's Wheel of Emotions [17], in addition to collecting multiple labels, we also use a probabilistic model to predict the discrete emotional distribution of the images.

2.2 Affective image annotation

There are two main approaches for collecting emotion labels for images. One is manual labeling [6, 12, 13], which is either done



(a). Interface of selection: selecting and displaying the candidate emotion words fairly; a player evaluates the emotion words in the descending order of emotion degree by dragging a line.

(b). Estimation mechanism: estimating the emotion degrees of the 24 emotion words based on the partial rank lists provided by multiple players.

(c). Incentive mechanism 1: game point and consistency degree with the opponent on each image in one game round.

(d). Incentive mechanism 2: consistency degree with all past players on eight primary emotions.

Figure 1: Overview of AffectI

offline or through crowdsourcing. The other is to utilize the text information associated with the images [2, 26, 27, 30]. While the text-based approach is suitable for constructing large-scale datasets, the quality of the data is usually low since the text information may not actually be related to the emotional perception of the images. Various natural language processing technologies have been employed to improve the quality of the collected data [30]. However, almost all of these automatically collected data need to be manually post-processed finally. FI dataset [27] was obtained through filtering 90,000 noisy images collected from Flicker and Instagram, and 225 AMT workers performed the filtering. Although some qualification tests to filter workers who just want to get paid for completing the task and do not perform the task seriously have been designed, fatigue due to the repeated labeling tasks can cause stress to workers, which may greatly affect their emotional perception of the images [24]. In contrast, by using game, our method enable players to evaluate images in a more relaxing and enjoyable way, and hence more reliable results can be expected.

2.3 Game with A Purpose

GWAP [22] is a human-based computation technique aimed at collecting data as a side effect of game play. The ESP game [21] is the classic online game for collecting labels for images. When an image is presented, two players input the words they think best represent the image. They will be rewarded with game points if they input the same word. To receive more points, players tend to give easier and more generic descriptions, and therefore the diversity of the output can be limited. To solve this problem, KissKissBan [10] introduced a third player, who acts as a blocker by setting blocking words to encourage more diverse labels. Karido [19] is a cooperative online game aimed at collecting descriptive labels. The descriptor input the text describing the target image, and the guesser selects the image based on the description. Artigo [4] is another game that proposed two methods of "square" and "script" to modify ESP games to collect deep semantic labels of image. All these existing games, however, can only collect descriptions of the object in the image, such as a bag, a car, a person, or some scenery represented by the image. To the best of our knowledge, no game is currently available for collecting labels describing the emotions evoked by images as well as the preferences of emotions. Existing GWAP implementations usually use the game points based on the consensus of players as the incentive [10, 19, 21, 23]. Such an incentive may encourage the players to guess the choices of opponents to achieve high scores instead of choosing the labels representing their own subjective emotional perception. AffectI is equipped with novel selection and incentive mechanisms enabling the collection of diverse labels.

3 OUR PROPOSAL

We propose an online game for affective image annotation, AffectI, based on the concept of GWAP, for collecting the emotional labels from images. "I" represents both Image and "I (myself)", which means it can identify the affects brought by images as well as the subjective emotion perceptions of the players.

The novelty of AffectI relies on three key mechanisms originally designed for addressing the subjective and relative perception issues of emotion. The first is the *selection mechanism*, which ensures the all emotion words being fairly evaluated and assists the players to select and judge the emotion degrees of words in an efficient way. The second is the *estimation mechanism*, i.e., which adapts the Bradley-Terry model [3] to estimate the emotion degree of all labels from partial rank list proved by multiple players. The third is the *incentive mechanism*, which shows the comparison between current player and her opponent or all past players and can encourage the player to provide labels with high quality.

3.1 Framework

Figure 1 shows the overview of AffectI. A player is advised to start with a step-by-step tutorial if she is not familiar with AffectI. The player can choose her opponent from a list of online players, who are either real online players or virtual players simulated from the records of past players. Therefore, a player can play AffectI at any time regardless of whether or not any online players are available. Figure 1.(a) is the selection screen, where an image to be labeled is displayed on the left, while the evaluation area is displayed in the middle, where six words from Plutchik's Wheel of Emotions are presented. The player needs to click on the word that best matches the emotion of the image and then drag the mouse to other words in the descending order of emotion degree. The player only needs to select the words that match the emotion of the image to a certain degree. If there are no matching words, she can select none of the words. After the current image is labeled successfully, a score computed based on the consistency with the opponent will be displayed on the top of the evaluation area, with words of praise if a high score has been achieved. One game round includes five images, and the player only has 20 seconds to evaluate each image. When one game round is over, the total score will be displayed, and the consistency of the emotion perception with that of the opponent (Figure 1.(c)) as well as that with all past players (Figure 1.(d)) will be visualized to enhance the player's experience and encourage her to play more rounds of games.

3.2 Selection Mechanism

If free word choice is allowed for annotating the images, it is unlikely that the players will select the same words. We thus collect the labels by asking players to select from the candidate emotion words. AffectI uses the 24 emotion words from Plutchik's Wheel of Emotions [17], which is designed to help users understand the nuances of emotion and the contrast between them. It has eight primary emotions, anger, disgust, fear, sadness, anticipation, joy, surprise, and trust, and each primary emotion has three strength levels (from intense to mild).

To obtain the emotion degree of each word for an image, a naïve solution is to let a player evaluate each word using common rating scales, such as Semantic Differential or Likert Scales. However, it is not easy for the players to assign an absolute degree. An alternative is to use pairwise comparison, which makes it easier for players to make the decisions. However, the total number of pairwise comparisons of 24 words is quite large. To reduce the number of comparisons for a player, we thus adopt a mechanism to make the players compare a subset of emotion words and give a partial ranking list based on the degree of emotional matching with the image. Figure 1. (a) shows an example of the players' selection interface. After a number of players have labeled the same image, we will then aggregate these sub rank lists into one whole rank list by estimating the emotion degrees. The estimation mechanism will be introduced in the next section.

For the subset of emotion words shown to the players, two issues need to be solved. One is how to select the subset; the other is how to show them. For the first issue, we need to determine the size of the subset. According to the Magic number 7 rule (Miller's law) [15] for interaction design, which says that the choice should be limited to 7 ± 2, we display six words in each trial. We design the following strategy for selecting the six candidate words from all 24 emotional words to ensure that all words can receive a fair opportunity to be selected.

On one hand, the words with higher probability of being a label for the image need to be suggested to facilitate the players' selection of higher-quality words (i.e., exploitation); on the other hand, the words that have been shown with a low frequency need to be evaluated further to obtain a more accurate estimation of the emotion degree (i.e., exploration). Therefore, our selection mechanism is a hybrid strategy that makes a trade-off between exploitation and exploration. Although the techniques proposed for the multi-armed bandit problem, such as the Upper Confidence Bound (UCB) method [1], can be somewhat useful here, our purpose is to estimate the emotion degree for all words rather than to find the one best word. With the UCB method, it is difficult to control the detailed proportion of exploitation and exploration. We thus utilize the following specific selection rule. In the six words shown to a player, we randomly select two words with a probability of being labeled that is higher than a given threshold θ, which is empirically set to 0.65. This labeled probability for an emotion word to a given image is defined as the ratio of the number of times the word is labeled by players (regardless of the rank in the selection) to the total number of times this emotion word is shown. We also select the four words with the lowest frequency of being shown to the players.

One of the crucial advantages of the proposed selection mechanism is that it encourages the players to annotate diverse emotions and to provide words with diverse emotion degrees in a primary emotion. For example, let us consider the case with the primary emotion "joy", the intense motion "ecstasy" and the mild emotion "serenity". If a player is asked to provide free words or select words from the 24 emotion words, there is a high probability that she will only select the word "joy". In contrast, in our system, because the other two words "ecstasy" and "serenity" will also be shown to the players with a fair frequency, these two words have a higher probability of being selected than the probability in other systems. Assuming that an image contains the primary emotion "joy", if the words "ecstasy" or "serenity" are shown without the word "joy", the player will select them; if the words "ecstasy" or "serenity" are shown with the word "joy", the player will select them and rank them based on the emotion degree. In summary, the chance that these two words will be selected is improved considerably. In contrast to only utilizing high frequency emotion words like [10], leveraging diverse emotion words also potentially improves the reliability of the emotional labels because it tends to represent the affective information of the images more accurately.

3.3 Estimation Mechanism

The images shown to the players are randomly selected from the image dataset. The same image will be evaluated with different sets of six emotion words multiple times. As mentioned above, each player only evaluate a subset of the emotion words. Thus, we need a method to estimate the emotion degree for all 24 emotion words by aggregating these partial rank lists. We employ the BT (Bradley-Terry) model [3], which can be used to estimate the ranking scores of all objects (emotion degree of all words) from partial pairwise comparison labels. There are variants of BT such as CrowdBT[5] have been proposed. The variants have different assumptions on modeling the inconsistent options of multiple players on same object. CrowdBT assumed that some players make mistakes because of their ability when there are inconsistent labels. In our work, we assume that inconsistent labels of multiple players are because of the personal perceptions. Therefore, we adapt the BT model to follow this assumption. Figure 1. (b) visualizes an example of the estimated degrees of the 24 emotions for an image based on the partial rank lists from multiple players.

For a given image a_k, we have a set of players $\mathcal{B}_k = \{b_x\}_x$ who have labeled this image. For a player b_x, the six words shown to this player for this image is determined by $\mathcal{W}_x^k = \{w_{i_u}\}_{u=1}^6$. Because we estimate the rank list for each image independently, we omit the subscript k representing the image in the following formulation, e.g., $\mathcal{B} = \mathcal{B}_k, \mathcal{W}_x = \mathcal{W}_x^k$. Assuming that the player selects m words from the six words with an order and drops the others, we can define the obtained partial ranking list as $l_x = \{w_{i_1} \succ_x \cdots \succ_x w_{i_m} \succ_x w_{i_{(m+1)}} \cdots w_{i_6}\}$. The dropped words have the same and lowest rank; any selected word has a higher rank than a dropped word. By traversing this ranking list, we can convert it to a set of pairwise preference comparisons with all the combinations of words in this list except the pairs of dropped words, i.e., $C_x = \{(w_{i_u} \succ_x w_{i_v}) | w_{i_u} \in \mathcal{W}_x, w_{i_v} \in \mathcal{W}_x, i_u \leq i_m\}$. The pairwise preference comparison by all players for the image a_k is defined as $C = \{C_x\}_x$.

The problem of the emotion degree estimation can be defined as follows. Given the player set $\mathcal{B}$, emotion word set $\mathcal{W}$, and pairwise preference comparison set C, we estimate the rank list $\hat{l}$ with the emotion degrees (rank scores) $\mathcal{S} = \{s_i\}_i$ of all emotion words in $\mathcal{W}$, where s_i is defined as the emotion degree of word w_i.

The detailed algorithm based on the BT model is described as follows. The probability $p(w_i \succ w_j)$ that a word w_i precedes a word w_j can be defined as

$$p_{ij} = p(w_i \succ w_j) = \frac{e^{s_i}}{e^{s_i} + e^{s_j}} = \frac{1}{1 + e^{-(s_i - s_j)}}. \tag{1}$$

From the collected pairwise preference comparison set C, we can compute the number of players that prefer w_i to w_j, i.e., n_{ij}. We use this to compute the observed probability q_{ij} that word w_i precedes word w_j, i.e., $q_{ij} = n_{ij}/(n_{ij} + n_{ji})$. Using the observed probability, we can estimate the emotion degree by minimizing the following objective function:

$$\begin{aligned} \mathcal{L} &= -\sum_{ij} \log \left(q_{ij} p_{ij} + (1 - q_{ij})(1 - p_{ij}) \right) \\ &= -\sum_{ij} \log \left(q_{ij} \frac{1}{1 + e^{-(s_i - s_j)}} + (1 - q_{ij}) \frac{1}{1 + e^{-(s_j - s_i)}} \right). \end{aligned} \tag{2}$$

The estimation mechanism estimates the emotion distribution of image from partial judgements. It allows the players to quickly evaluate small subsets of the emotion words, reduces the time costs of players and thus is efficient. It provides a solution to aggregate the perceptions from multiple players and thus the aggregated emotional labels have higher reliability.

3.4 Incentive Mechanism

In AffectI, we propose two different kinds of incentives. One is based on a comparison between the current player and her opponent, which encourages the player to provide high-quality labels. The other is based on a comparison between the current player and all past players. From the empirical results, we find that more players tend to provide diverse labels which are different from that provided by the majority after applying this incentive.

3.4.1 Opponent-based Incentive (OI).

We show the comparisons between a player and her opponent in two ways. One is through game points, as in existing works [10, 19, 21, 23], and the other is a visualization of the degree of consistency. They encourage the players to provide high-quality labels, i.e., the accurate labels that are related to the given images.

Given the partial ranking list l_x and l_y of two opponent of the players b_x and b_y to a specific image a_k, we denote r_{xi} as the rank of word w_i in rank list l_x. Note that all unselected words have the same rank. For each candidate word, a player can obtain a high amount of points if she assigns the same rank to a word as her opponent, i.e., $o_{xi} = 2\tau$, if $r_{xi} = r_{yi}$, or she can receive a medium number of points if she assigns a similar rank, i.e., $o_{xi} = \tau$, if $r_{xi} = r_{yi} \pm 1$. τ is an integer value controlling the scale of the points, such as $\tau = 5$.

We also emphasize the importance of top-ranked words, as they have a higher probability of being used as the annotation of the given image. For details, $o_{xi} = 8\tau$, if $r_{xi} = r_{yi} = 1$; $o_{xi} = 6\tau$, if $r_{xi} = r_{yi} = 2$; $o_{xi} = 4\tau$, if $r_{xi} \in \{1, 2\}$ and $r_{yi} = 2 - r_{xi}$. Finally, the total points that a player obtains is calculated by $o_x = \sum_i o_{xi}$.

We visualize the consistency of the selection order of emotion words between a player and the opponent. It is computed by the Spearman Rank Correlation Coefficient ρ, which is defined as $\rho_k = 1 - (6 \sum (r_{xi} - r_{yi})^2)/(n^3 - n)$, where $n = 6$ is the number of words shown to the two players. For each image a_k evaluated, we compute ρ_k. $\rho = \sum_k \rho_k$ for all five images are presented at the end of each game round. Figure 1.(c) shows an example.

3.4.2 Past Players-based Incentive (PPI).

We also visualize the matching of emotion perception of a player with that of all past players. This is to enable the current player to understand which emotions she is reporting consistently with other players and which she is not. Plutchik's Wheel of Emotion model has eight primary emotions defined by $C = \{c_z\}_z$. Based on the annotations of the five images in one game round, we compute the consistency degree of a player with all past players on each primary emotion.

For an evaluated image a_k, we have the player's partial rank list l_x for the six candidate words and the estimated full rank list $\hat{l}$ of all 24 words computed by the estimation mechanism described in Section 3.3 from the annotations of all past players. We extract the order of the six words from full rank list $\hat{l}$ and generate the estimated partial rank list $\tilde{l}$ only containing these words; the corresponding rank of word w_i in rank list $\tilde{l}$ is defined as $\tilde{r}_i$. Then, the consistency degree for a word w_i between the selection of the current player and that of all past players is computed by $\phi_{xi}^{k} = 1/(1 + |r_{xi}^{k} - \tilde{r_i}^{k}|)$.

Each primary emotion has three strength levels, intense, normal and mild, and each level in a primary emotion corresponds to an emotion word. We thus assign a weight $\gamma_i \in \{1/2, 1/3, 1/6\}$ based on the strength levels to each word w_i. The overall consistency degree between a player b_x and all past players on a primary emotion c_z is computed as follows: $\alpha_{xz} = \sum_{w_i \in c_z} \sum_{k=1}^{5} \gamma_i \phi_{xi}^{k} / 5|c_z|$, where $\phi_{xi}^{k} = 0$ if w_i is not shown in the six words for image a_k. Figure 1.(d) illustrates an example. We empirically investigate the influence of this incentive in the experimental section and find that it can encourage players to provide diverse labels.

4 EXPERIMENTS

4.1 Implementation

AffecI was implemented using client-server-client architecture. The server is built with node.js and socket.io is used for communication between clients. SQLite is used for data collection and storage. The design of the interface mainly relies on HTML, CSS, and jQuery, combined with processing.js for the visual display of images and words. A sliding gesture interface is used for the selection of labels for the easy operation on cell phones though AffectI can also be played on desktop and laptop computers. AffectI keeps the history of past players and therefore, as introduced in Section 3.1, a player can either choose a current online player or a virtual player simulating a past player as the opponent.

4.2 Dataset

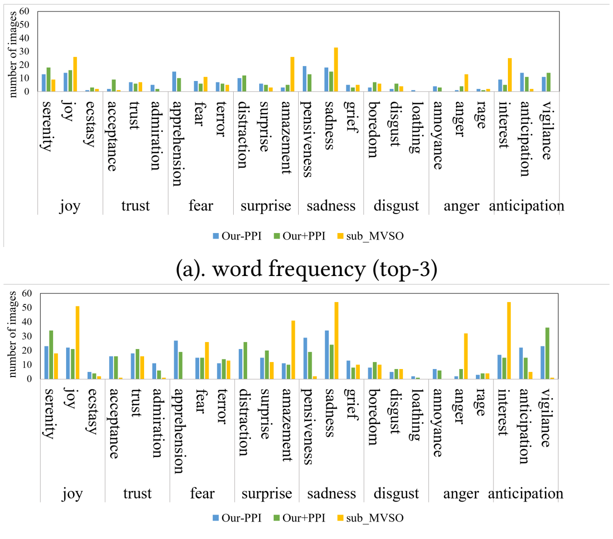
For comparison, we selected 60 images from the MVSO dataset [11], which is a multilingual version of FlickerCC dataset[2] constructed using Adjective-Noun Pairs (ANP). ANPs, such as “beautiful flower” or “sad eyes”, were retrieved by using emotional words from Plutchik’s wheel of emotions and then selected based on the consideration of frequency and diversity. There are more than 4,000 ANPs in MVSO with each ANP associated with multiple images. In our experiment, seven to nine images with high web views were selected for each of the 8 primary emotion categories. Each ANP has the scores for the 24 emotions, and one image may be associated to multiple ANPs. We used the average emotion scores of all ANPs associated to an image as the emotion score. In this way, we obtain a discrete distribution of the 24 emotions for each of the 60 images.

4.3 Experimental Design

We designed ab experiment to investigate the following questions.

- **Q1. Diversity:** Can the system successfully collect diverse emotional labels with different emotion degrees compared to the existing work?
- **Q2. Reliability:** What is the quality of the diverse labels collected by our system?
- **Q3. User Experience:** What is the player experience when using our system?

For the issue of efficiency of our system on collecting the labels, because our selection and estimation mechanisms naturally allow the players to quickly evaluate small subsets of the emotion words and provide the partial rank list. It can naturally reduce the time costs of players and thus is efficient.



(a). word frequency (top-3)

(b). word frequency (top-6)

Figure 2: Diversity evaluation by the frequency of annotated emotion words

We conducted the experiment with our system. The participants age from 22 to 35. We respectively collected the emotional labels without or with the mechanism of Past Players based Incentive (PPI, Section 3.4.2). For the games carried out without PPI, we name this setting of our system as "Our-PPI". There are 163 unique player IDs. The 60 images are labeled 1,892 times in total. From all $1,892 \times 6 = 11,352$ emotion words shown in the games, there are 4,479 emotion words selected to annotate the images. For the games carried out with PPI, we name this setting of our system as "Our+PPI". There are 67 unique player IDs. The 60 images are labeled 710 times in total. From all $710 \times 6 = 4,260$ emotion words shown in the games, there are 1,546 labels selected to annotate the images. The annotations in the subset of the MVSO dataset are used as the baseline annotations.

4.4 Q1. Diversity

We compare the diversity of emotional labels collected by our system and MVSO. First, we evaluate the diversity by calculating the frequency of the emotion words annotated in the top-k words of 60 images for each system. It is to evaluate that whether all the 24 emotion words are appropriately used to represent the different degree of emotions for all the images in the dataset. Figure 2.(a) shows the result in the top-3 case and Figure 2.(b) shows the result in the top-6 case. We group the emotion words based on the primary emotions.

The results show that sub_MVSO mainly only use the representative word of each primary emotion and rarely use the intense and mild emotion words. For example, for the "joy" primary emotion, the word "joy" has much higher frequency than the intense emotion word "ecstasy" and the mild emotion word "serenity" in both

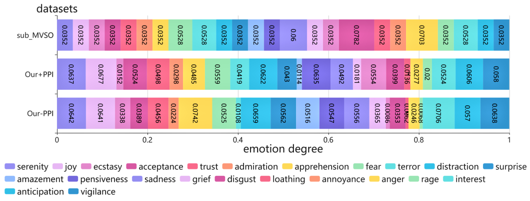


Figure 3: Emotion degrees of the emotion words to an image

Figure 2.(a) and 2.(b). Figure 2.(b) also shows that sub_MVSO almost assigns the "joy" emotion word to 50 of 60 images in the top-6 annotations. In other words, MVSO mainly use the high frequency emotion words in the corpus and lack of diversity.

In contrast, for both "Our-PPI" and "Our+PPI", the frequencies of the emotion words in a primary emotion are relative uniform. The intense and mild emotion words have more chances to be annotated to the images. It shows that the emotion labels collected by our system have higher diversity than those of MVSO.

Second, we evaluate the diversity on the estimated emotion degrees of the emotion words. Figure 3 gives an example on this type of diversity. It shows the distribution of the emotion degrees for an image randomly selected from the 60 images. It shows that for both Our-PPI and Our+PPI, there are obvious differences among the emotion degrees of different emotion words for the target image. However, for sub_MVSO, except for a few typical emotion words, other emotion words almost have the same emotion degrees even though they are from different primary emotions.

It is because that MVSO calculates the emotion scores of ANP based on the number of times that the images are associated with both ANP and emotion keywords. There is no comparisons among the emotions. The estimated emotion degrees in MVSO thus lack of diversity among the emotion words. Our system estimates the emotion degrees based on the comparisons of emotion words. It thus can assign diverse emotion degrees to the emotion words and represent the relative degrees of the words better.

We also quantitatively evaluate the diversity on the estimated emotion degrees of the emotion words by using an entropy-based measurement. For an image a_k, the entropy of the emotion degrees of all emotion words is computed by $h_k = -\sum_{i=1}^{24} s_i \log s_i$, where s_i is the emotion degree of word w_i. When the entropy value is high, it means that the emotional information is more dispersed and the emotional labels are more diverse. Figure 4 shows the distribution of the entropy on all images for each system. It shows that both Our-PPI and Our+PPI have higher entropy than sub_MVSO. $p < 0.05$ in the t-test on the results of each pair of the systems. The results are statistically significant. The emotion labels collected by our system have higher diversity on the emotional degrees.

In addition, Our+PPI has higher entropy than Our-PPI, which shows the influence of the PPI incentive on the diversity. With the PPI incentive, more players tend to choose the labels representing their own subjective emotional perception. The potential reason may be that more players are curious about how their feelings differ from the other individuals in the majority.

As the summary of the diversity evaluation, based on the comparison between Our-PPI and sub_MVSO, as discussed in Section 3.3, the selection mechanism increases the diversity of emotional labels by enabling the players to annotate diverse emotions; based on the comparison of Our-PPI and Our+PPI, we find that, in addition to the selection mechanism, adding the PPI incentive mechanism is also beneficial for increasing the diversity of emotional labels.

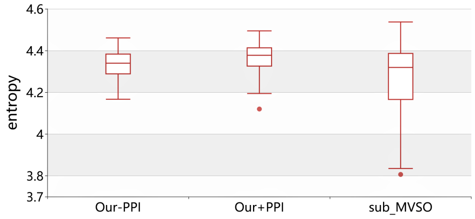


Figure 4: The distribution of the entropy of each method on the 24 emotions for each image.

4.5 Q2. Reliability

We have verified that our system can collect diverse labels with different emotion degrees. We need to verify the reliability of these diverse labels. Because evaluating the annotations for the images in the entire dataset is difficult, we extracted 10 images which have largest differences between the annotations of our system and sub_MVSO for human evaluation on the label reliability. These 10 images were divided into three groups; each group was evaluated anonymously by seven evaluators who had not participated in the AffectI game. We ask the evaluators to provide inaccuracy, neutrality, and accuracy judgments to each emotion word in the top-6 emotion words on the basis of the emotion degrees obtained by our systems and sub_MVSO.

Figure 5 illustrates the average accuracy and inaccuracy of the emotion words in the top-k ($k \leq 6$) emotional labels. For the emotion words in top-k ($k \leq 6$) emotional labels, the accuracy of our systems are always higher, and the inaccuracy of our systems are always lower than MVSO. MVSO mainly uses the high frequency emotion words in the corpus which not only results in the lack of quality, but also the lack of reliability. The emotional labels collected by our systems can describe the affective information of the images more accurately. Although it has shown that the labels diversity of Our+PPI is high in Section 4.4, the label reliability of Our+PPI is lower than Our-PPI. One of possible reasons is that Our+PPI encourages the players to provide more personal emotion labels which are diverse but may be inconsistent with the options of the majority. Our+PPI still collects the labels based on the subjective perceptions of players rather than only use the high frequency words and thus outperforms MVSO.

We show three detailed examples in Figure 6. In the example of Figure 6.(a), the affective information of the image includes "anticipation" which is annotated by our system. On the contrary, the emotion words such as "sadness" and "amazement" that are annotated by MVSO are not prominent in the image. In the example of Figure 6.(b), the emotional labels collected by our system such as "apprehension" can better describe the affective information in this image. On the contrary, the emotion labels annotated by MVSO,

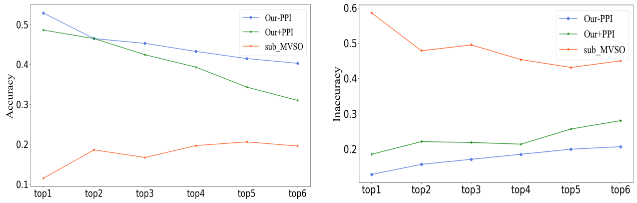
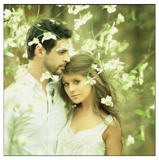


Figure 5: The average accuracy and inaccuracy of the emotion words in the top-k ($k \leq 6$) emotional labels by our systems and sub_MVSO.

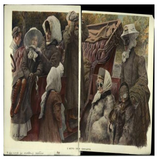


Our-PPI	In.	N.	Acc.
anticipation	0/7	1/7	6/7
interest	0/7	4/7	3/7
joy	0/7	2/7	5/7
sub_MVSO	**In.**	**N.**	**Acc.**
sadness	6/7	1/7	0/7
amazement	4/7	2/7	1/7
joy	0/7	2/7	5/7

(a). Example 1

Our-PPI	In.	N.	Acc.
apprehension	1/7	1/7	5/7
pensiveness	1/7	3/7	3/7
grief	2/7	4/7	1/7
sub_MVSO	**In.**	**N.**	**Acc.**
ecstasy	6/7	1/7	0/7
boredom	3/7	3/7	1/7
joy	6/7	1/7	0/7

(b). Example 2



Our-PPI	In.	N.	Acc.
terror	1/7	5/7	1/7
apprehension	2/7	4/7	1/7
pensiveness	1/7	1/7	5/7
sub_MVSO	**In.**	**N.**	**Acc.**
acceptance	4/7	3/7	0/7
interest	1/7	2/7	4/7
joy	4/7	2/7	1/7

(c). Example 3

Figure 6: Examples of Affective Image Annotation. In.: Inaccuracy; N.: Neutrality; Acc.: Accuracy.

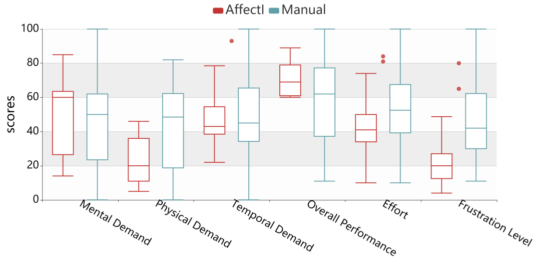


Figure 7: NASA-TLX evaluation results

such as "ecstasy" and "joy" are not reflected in the image. In the example of Figure 6.(c), the affective information of the image includes "pensiveness" which is annotated by our system. On the contrary, the emotion words such as "acceptance" and "joy" that are annotated by MVSO are not prominent in the image.

Furthermore, for the examples in Figure 6, MVSO always uses the normal emotion words with high frequency in the corpus, such as "joy" and "sadness"; in contrast, our system can utilize diverse emotion words to annotate the image. It verifies that our system can collect diverse and reliable emotional labels.

4.6 Q3. User Experience

To evaluate the user experience, We conducted a manual labeling experiment involving 28 participants for comparison. The 60 images were divided into 4 groups and each participant was asked to label one group of images (15 images) manually. They were asked to select the matching words from all of the 24 emotion words for each of the 15 images. After the experiment, 15 players of AffectI and the 28 participants of the manual evaluation experiment were asked to fill NASA Task Load Index (NASA-TLX [5]) form. The results are shown in Figure 7. We can find that AffectI outperforms the manual evaluation for physical demand, performance, effort, and frustration level. Especially, there is a significant improvement in frustration level, which demonstrated the advantage unique to a game. On the other hand, the result shows that AffectI has a slightly higher mental demand. In the post-experiment interview, we found such mental depend is related to the temporal demand, and the players of AffectI commented that they needed to be quite concentrate on the task given the time limit of 20 seconds for evaluating each image. Being aware of the existence of opponent seems also having added some mental load to the players. We did not set any time limit for the manual evaluation and the major temporal demand of manual labeling seems due to the pressure to select from all 24 words. Almost all participants of manual evaluation commented that it was very tough and time consuming.

From the results of NASA Task Load Index as well as the participants' comments, we can conclude that by using GAWP, the proposed method could improve user experience, and, especially, enable them to label the images in a less frustrating way. This might have contributed to the reliability of the collected labels mentioned in Section 4.5.

5 CONCLUSION

In this paper, we proposed a novel affective image annotation system, AffectI, which could efficiently collect high-quality and diverse emotional labels together with the emotion degrees of these labels for images. The selection mechanism makes all candidate emotion words have fair opportunities to show and contributes the diversity and quality of the labels; the estimation mechanism assessed the emotion distribution of image from partial judgements and thus contributed to the reliability and efficiency. The incentive mechanism encouraged the player to provide labels with high quality. We also found that the past-player-based incentive also promote the players to provide diverse labels. We constructed a novel image dataset with emotion distribution resulting from our experiment which can be used for affective image content analysis. In the future work, we will enlarge the dataset and estimate the personalized emotion degrees of players for images.

ACKNOWLEDGMENTS

This work was partially supported by JSPS KAKENHI Grant Number 17H00738, 19H05472 and 19K20277; the National Science Fund of China Grant Number 61871170.

REFERENCES

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002), 235–256.

[2] 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 *Proceedings of the 21st ACM international conference on Multimedia*. 223–232.

[3] Ralph Allan Bradley and Milton E. Terry. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345. http://www.jstor.org/stable/2334029

[4] François Bry and Christoph Wieser. 2012. Squaring and scripting the esp game: Trimming a gwap to deep semantics. In *International Conference on Serious Games Development and Applications*. Springer, 183–192.

[5] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 193–202.

[6] Elise S Dan-Glauser and Klaus R Scherer. 2011. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior research methods* 43, 2 (2011), 468.

[7] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[8] Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1734–1748.

[9] Alan Hanjalic. 2006. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine* 23, 2 (2006), 90–100.

[10] Chien-Ju Ho, Tao-Hsuan Chang, Jong-Chuan Lee, Jane Yung-jen Hsu, and Kuan-Ta Chen. 2009. KissKissBan: a competitive human computation game for image annotation. In *Proceedings of the acm sigkdd workshop on human computation*. 11–14.

[11] 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.

[12] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* 1 (1997), 39–58.

[13] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*. 83–92.

[14] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. 2005. Emotional category data on images from the International Affective Picture System. *Behavior research methods* 37, 4 (2005), 626–630.

[15] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[16] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 860–868.

[17] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*. Elsevier, 3–33.

[18] Harold Schlosberg. 1954. Three dimensions of emotion. *Psychological review* 61, 2 (1954), 81.

[19] Bartholomäus Steinmayr, Christoph Wieser, Fabian Kneißl, and Fracois Bry. 2011. Karido: A GWAP for telling artworks apart. In *2011 16th International Conference on Computer Games (CGAMES)*. IEEE, 193–200.

[20] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)* 3, 3 (2007), 1–13.

[21] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 319–326.

[22] Luis Von Ahn, Shiry Ginosar, Mihir Kedia, Ruoran Liu, and Manuel Blum. 2006. Improving accessibility of the web with a computer game. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 79–82.

[23] Luis Von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 55–64.

[24] Kyosuke Watanabe, Akihiro T Sasaki, Kanako Tajima, Kenji Mizuseki, Kei Mizuno, and Yasuyoshi Watanabe. 2019. Mental fatigue is linked with attentional bias for sad stimuli. *Scientific reports* 9, 1 (2019), 1–8.

[25] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. 2017. Learning visual sentiment distributions via augmented conditional probability neural network. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[26] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Qicong Chen, Juanzi Li, Chunxiao Xing, and Jie Tang. 2014. How do your friends on social media disclose your emotions?. In *Twenty-eighth AAAI conference on artificial intelligence*.

[27] 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 *Thirtieth AAAI conference on artificial intelligence*.

[28] Sicheng Zhao, Guiguang Ding, Yue Gao, and Jungong Han. 2017. Approximating discrete probability distribution of image emotions by multi-modal features fusion. *Transfer* 1000, 1 (2017), 4669–4675.

[29] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. 2018. Affective Image Content Analysis: A Comprehensive Survey. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (Stockholm, Sweden) *(IJCAI'18)*. AAAI Press, 5534–5541.

[30] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. 2016. Predicting personalized emotion perceptions of social images. In *Proceedings of the 24th ACM international conference on Multimedia*. 1385–1394.