Median based Multi-label Prediction by Inflating Emotions with Dyads for Visual Sentiment Analysis

Tetsuya Asakawa * and Masaki Aono†
*† Toyohashi University of Technology, Aichi, Japan
* E-mail: asakawa@kde.cs.tut.ac.jp
† E-mail: aono@tut.ac.jp

***Abstract*—Visual sentiment analysis investigates sentiment estimation from images and has been an interesting and challenging research problem. Most studies have focused on estimating a few specific sentiments and their intensities. Multi-label sentiment estimation from images has not been sufficiently investigated. The purpose of this research is to accurately estimate the sentiments as a multi-label multi-class problem from given images that evoke multiple different emotions simultaneously. We first introduce the emotion inflation method from six emotions defined by the Emotion6 dataset into 13 emotions (which we call 'Transf13') by means of emotional dyads. We then perform multi-label sentiment analysis using the emotion-inflated dataset, where we propose a combined deep neural network model which enables inputs to come from both hand-crafted features (e.g. BoVW (Bag of Visual Words) features) and CNN features. We also introduce a median-based multi-label prediction algorithm, in which we assume that each emotion has a probability distribution. In other words, after training of our deep neural network, we predict the existence of an evoked emotion for a given unknown image if the intensity of the emotion is larger than the median of the corresponding emotion. Experimental results demonstrate that our model outperforms existing state-of-the-art algorithms in terms of subset accuracy.**

I. Introduction

With the spread of SNS and the Internet, a vast number of images are widely spread. As a result, there is an urgent requirement for image indexing and retrieval techniques. In fact, we can feel several emotions from an image. Different visual images have different emotional triggers. For instance, an image with a snake or a spider may most likely trigger a bad feeling like "disgust" or "fear," whereas an image with a flower may most likely trigger a good feeling like "amusement" or "excitement".

Most previous studies have focused on estimating a few specific sentiments, and the multi-label sentiment (joy, sadness, anger, excitement, surprise, fear) estimation from the images has not been sufficiently investigated. Furthermore, visual sentiment prediction investigates sentiment estimation from images and has been an interesting and challenging research problem. The purpose of this research is to accurately estimate the sentiments as a multi-label multi-class problem from given images that evoke multiple different emotions simultaneously.

Fan et al. [1] performed sentiment prediction using the Emotion6 dataset. However, the existing Emotion6 dataset has small number of items, and it is difficult to predict multiple emotions. Therefore, we first describe a dataset that simulates Plutchik's wheel of emotions, which can be constructed from the Emotion6 dataset. We describe 'Transf13' in a dataset. No research to our knowledge has used Transf13. We then perform a multi-label sentiment analysis using the dataset.

In this paper, we also employ a Color Histogram (CH) as another hand-crafted feature, which is a representation of the distribution of colors in an image which results in hand-crafted features. In the meantime, we also use a BoVW method as a representative method for extracting hand-crafted features of images. To implement a BoVW model, here we adopt an ORB [2] as a local feature. We study to combine several hand-crafted features for visual sentiment analysis. We also introduce a new combined neural network model which allows inputs coming from both hand-crafted features such as BoVW (Bag of Visual Words) and pre-trained CNN features. In addition, existing deep learning had weak classifications, therefore we propose a new fully connected 2 layers. The new contributions of this paper include; (1) propose a novel feature considering both hand-crafted and CNN features to predict sentiment of images, unlike most recent research only concerns adopting hand-crafted or CNN features, (2) propose a combined feature method to combine the output of each feature, unlike previous work which focuses on combining feature vectors.

In the following, we first survey related work in Section II, followed by introducing three features used in this research in Section III. Section IV is introducing the dataset and experimental settings. In Section V, we describe experiments we have carried out, and conclude this paper in Section VI.

II. Related work

Visual sentiment analysis is an important task which has seen rapid development in recent years. Research on sentiment analysis is roughly divided into two approaches in terms of how many specific sentiments should be classified. The first research group has dealt with "Positive," "Negative," and sometimes "Neutral," while the second research group has dealt with more minute categories, typically based on Plutchnik's Wheel of Emotions [3].

Examples of the first research group includes Solli et al [4], who performed emotion prediction using BoVW features, Katsurai et al [5], who exploited latent correlations among visual, textual, and sentiment features for image sentiment classification, Soleymani et al [6] surveyed multimodal

sentiment analysis, Xu et al [7], who transferred VGG networks trained on ImageNet dataset into visual sentiment analysis on the sentiment datasets,

who transferred VGG networks trained on the ImageNet dataset into visual sentiment analysis on the sentiment datasets, Fan el al. [8], who studied the relation between image sentiment and visual attention, and Cordel et al. [9], who proposed emotion-aware human attention prediction.

On the other hand, examples of the second research group include Kim et al [10], who used feedforward deep learning for 8 class visual sentiment classification, Liu et al [11], who presented multi-label visual sentiment distribution prediction model from images of 8 emotions, and Yang et al [12], who proposed weakly supervised coupled networks for visual sentiment analysis using 4 emotions ("Joy", "Fear", "Anger", and "Sadness").

On the other hand, examples of the second research group include Kim et al. [10], who used feed forward deep learning for 8-class visual sentiment classification, Liu et al. [11],

who presented multi-label visual sentiment distribution prediction model from images of 8 emotions, and Yang et al. [12], who proposed weakly supervised coupled networks for visual sentiment analysis using 4 emotions ("Joy," "Fear," "Anger," and "Sadness"). Microsoft Emotion API [13] also provides 8 emotions including "Happiness," "Angry," "Disgust," "Fear," "Contempt," "Sadness," "Surprise," and "Neutral".

To our knowledge, none of the above research has dealt with "multi-label" classification of emotions. We perform multi-label visual sentiment analysis which our proposed emotion with dyads based on previous research [14]. Borth et al [3] performed visual sentiment analysis from images based on Plutchik' s wheel of emotions. However, there was no research on performing multi-label visual sentiment analysis by inflating emotions with dyads. To evaluate multi-label multi-class classifications, we adopted exact match (or subset accuracy [15]) and the mean of the each class-average precision (mAP) [16].

III. PROPOSED METHOD

We propose a multi-label visual sentiment analysis system to predict multiple emotions inspired by Plutchnik's wheel of emotions. To this end, we start with six emotions provided by the Emotion6 dataset, followed by inflating emotions into 13 categories using dyads. After showing our emotion inflation method, we will describe our deep neural network model that enables the multi-label outputs, given images that evoke emotions.

A. Proposed emotion inflation method

Here we adopt Emotion6 [17] as a basis for our emotion inflation. Each image in the Emotion6 dataset has an emotion distribution indicating the probability of Ekman' s six basic emotions [18]. Emotion6 contains a total of 1,980 images and each image is labelled with six kinds of emotions, i.e. "Anger," "Disgust," "Fear," "Joy," "Sadness," and "Surprise."

We propose to inflate the six emotions mentioned as above hoping to capture abundant psychological human feelings. Specifically, we employ Plutchik's emotional dyads to expand Emotion6, where the dyads are supposed to represent two emotions occurring simultaneously. It is noted that, by simply combining Plutchik's 8 emotions [19] [20], we will acquire 24 combined emotional dyads. However, we have Emotion6 as our basis, which has 6 emotions instead of 8, so that we end up 13 emotional dyads out of 24. Hereafter we call the 13 dyads $Transf13$ as shown in Table I. An example of images corresponding to each emotion of Transf13 is depicted in Figure 1.

Each image provided by Emotion6 is associated with 6 emotion probabilities, instead of a single emotion label as ground truth. In our Transf13 inflated emotion model, we assume that we can extract two pairs of emotions independently from the emotion distribution of an arbitrary image. Mathematically, we hypothesize that the following equation holds.

$$\text{Prob}(O(Z)) = \text{Prob}(I(X))) + \text{Prob}(I(Y)), \qquad (1)$$

where I represents an input image, X and Y represents different emotions from Emotion6, O represents an output image, Z represents an emotion from Transf13.

Please note that given two different arbitrary emotions from Emotion6, the number of combinations is ${}_6\mathrm{C}_2$ or 15, while two dyads, i.e., "Anger" + "Fear," and "Joy" + "Sadness" never happen, so that we have 13 emotions in total, which we call **Transf13** hereafter. We repeat the probability computation defined above for every image in the Emotion6 dataset. Then, we propose to generate a multi-hot vector representing a new ground truth for each image, based on the 13 emotion distributions, where we introduce a threshold parameter.

Specifically, suppose we have an image with emotion distribution as shown in 2, we generate a multi-hot vector by specifying a certain threshold, which we call *Minimum Emotion Evocation Value* (MEEV). For example, if we set MEEV = 0.2, then we will generate a multi-hot vector such as $[0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1]$. The details will be described in the Experiment section. With the generated multi-hot vectors and the provided Emotion6 dataset, we reduce our problem to a multi-label, multi-class classification problem.

B. Proposed deep neural network model

To solve our multi-label, multi-class classification problem, we propose new combined neural network models which allow inputs coming from both hand-crafted and End-to-end (CNN) features. Hand-crafted features we have adopted include a Color Histogram (CH) feature, a taking care of global feature, and Bag of Visual Words (BoVW), taking care of a local feature. On the other hand, CNN features extracted from a pre-trained CNN-based neural network, include VGG16 and ResNet50.

In order to deal with the above combined features, we propose a deep neural network architecture where we allow multiple inputs and a multi-hot vector output.

Fig. 1. Image for each emotion of Transf13.

TABLE I
SELECTED THIRTEEN COMBINATIONS (TRANSF13).

Number	Transf13	Emotions A	Emotions B
1	Contempt	"Anger" + "Disgust"	
2	Pride	"Anger" + "Joy"	
3	Envy	"Anger" + "Sadness"	
4	Outrage	"Anger" + "Surprise"	
5	Shame	"Disgust" + "Fear"	
6	Morbidness	"Disgust" + "Joy"	
7	Remorse	"Disgust" + "Sadness"	
8	Unbelief	"Disgust" + "Surprise"	
9	Guilt	"Fear" + "Joy"	
10	Despair	"Fear" + "Sadness"	
11	Awe	"Fear" + "Surprise"	
12	Delight	"Joy" + "Surprise"	
13	Disappointment	"Sadness" + "Surprise"	

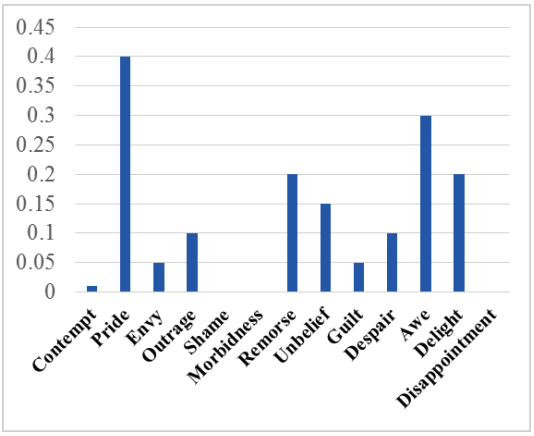

Fig. 2. An example of 13 emotion distribution

The combined feature is represented by the following formula:

$$\text{combined\_feature} = w_1(\text{CH}) + w_2(\text{BoVW}) + w_3(\text{CNN}) \quad (2)$$

Based on this formula, after the training process, we allow our neural network system predict the visual sentiment of unknown images as a multi-label multi-class classification problem. The overview of our proposed feature for feature extraction is shown in 3.

1) CH feature extraction: CH of an image represents the distribution of the composition of colors in the image.

CH of an image represents the distribution of colors in the image. According to Augello et al[21], colors and emotions are closely related. Thus we adopt CH as one of the features for our visual sentiment analysis system. Here we employ RGB color space for CH, where each color (i.e. Red, Green, Blue) has 2-bits, amounting to a 64 dimensional feature vector.

Although CH is a global and primitive feature, we incorporate CH into our system as one input as shown in Figure 3. Specifically, we prepare a fully connected (Dense) layer having a 64 dimensional vector as the input and a 256 dimensional vector as the output with "ReLU" for the activation function, where 256 dimension is empirically determined.

2) BoVW Feature extraction: "Bag of Visual Words" (BoVW), sometimes referred to as "Bag of Features," or "Bag of Keypoints" [22] have been used extensively in computer vision and image recognition as a feature representation, taking care of local features of an image. Since BoVW regards an image as an aggregation of local features, we need local features behind BoVW. In the following, we briefly describe the local feature we have adopted and the reason why, followed by how we incorporate BoVW into our system.

- **ORB (Oriented FAST and Rotated BRIEF)** Rublee et al [2] proposed ORB, Oriented FAST and Rotated BRIEF. Unlike SIFT and SURF [23], ORB is a license-free local feature taking advantage of the FAST keypoint factor [24] and the BRIEF descriptor [25].
  We have chosen ORB as a local feature descriptor for BoVW because ORB is fast, less prone to change in orientation, and it performs well under noisy conditions. In addition, ORB is known to be robust under conditions of lighting, blur, and perspective distortion, inherited from BRIEF.
- **Neural network for BoVW feature**
  We extract ORB features as visual words from all the training images given. Then, we apply a *k*-means algorithm to obtain 256 clusters for representative ORB features as visual words. The 256 clusters are then used for each image to obtain a histogram consisting of the frequencies of each bin, which can be regarded as a 256

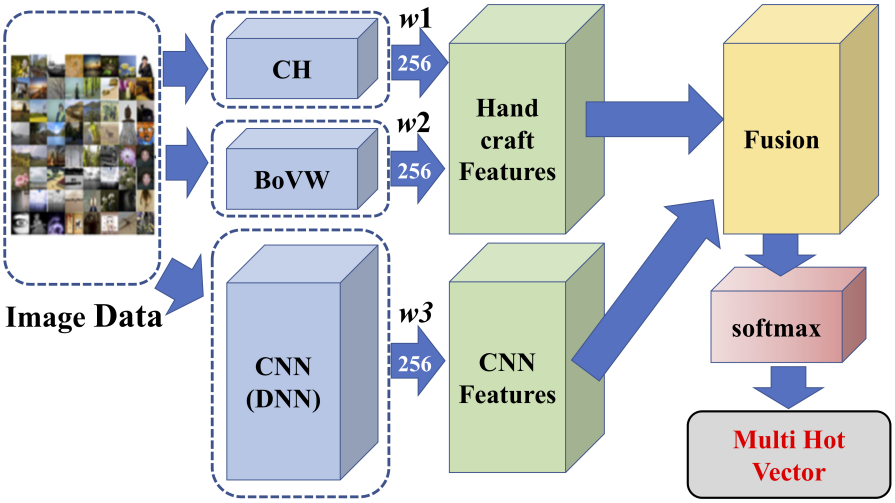

Fig. 3. Our proposed feature for multi-label feature extraction.

dimensional vector.

We incorporate BoVW into our system like CH as another input to our system as shown in Figure 3. To be specific, we prepare another fully connected (Dense) layer having a 256 dimensional vector as the input and the same 256 dimensional vector as the output with "ReLU" for the activation function.

3) CNN feature extraction: In addition to hand-crafted features described above, our system incorporates CNN features, which can be extracted from pre-trained deep convolutional neural networks with ImageNet [26] such as VGG16 [27], ResNet50[28], and NasNet-Large [29]. Because of the lack of dataset in visual sentiment analysis, we adopt transfer learning in our feature to prevent over fitting.

We decrease the dimensions of a fully-connected layers used in CNN models. Specifically, for VGG16, we extract a 4096 dimensional vector from 'fc2' layer (or the second to the last fully-connected layer), and reduce the vector to 256 dimension by applying a fully-connected layer. Similarly, for ResNet50, we extract a 2046 dimensional vector from 'avg_pool' layer (or GlobalAveragePooling2D layer), and reduce it to 256 dimension. For NasNet-Large, we extract a 4032 dimensional vector from 'avg' layer (or GlobalAveragePooling2D layer), and reduce it to 256 dimension. Note that the output of 256 dimension is determined empirically.

C. Combined Feature extraction

As illustrated in Figure 3, three features (CH, BoVW, and CNN features) are combined and represented by an integrated feature as linearly weighted sum, where weights are w_1 for CH, w_2 for BoVW, and w_3 for CNN features, respectively. As shown in Figure 4 Note that CH and BoVW features are combined first as hand-crafted features, and both hand-crafted and CNN features are passed out on "Fusion" processing to generate the integrated features, followed by "softmax" activation function.

D. Predicting multi-hot vector

To detect a multi-hot vector, we propose a method illustrated in Algorithm 1. The input is a collection of features extracted from each image with K kinds of sentiments, while the output is a K-dimensional multi-hot vector.

In Algorithm 1, we assume that the extracted features (here CH, BoVW, and CNN) are represented by their probabilities. For each sentiment (in our case, we have 13 sentiments in total with Transf13), we sum up the probabilities of the features, followed by averaging the result, which is denoted by T_i^k in Algorithm 1. It should be noted that the probability distribution for each sentiment might take different minimum and maximum. Thus, instead of using a fixed threshold for every sentiment, we have adopted "median" of each sentiment computed by the training data. For each feature, we employ the median as the threshold of the corresponding emotion evocation. After obtaining all the thresholds dynamically determined based on the medians, the multi-hot vector of each image is generated such that if T_i^k is equal to or greater than the average of all the thresholds, we set $S_i^k = 1$; otherwise $S_i^k = 0$, where S_i^k is the element of k-th sentiment of i-th image. In short,the vector $\mathbf{S_i}$ represents the output multi-hot vector. We repeat this computation until all the test (unknown) images are processed.

IV. EXPERIMENTS

Here we describe experiments and the evaluations. The dataset we have used is Emotion6 [1] as our basis as described before. We have converted Emotion6 dataset and inflated the six dimension to 13, which we call Transf13. The dataset consists of 1,980 images originally, where we reduce this data to 1,402 by means of MEEV mentioned in Section III. We have divided the reduced data into training and testing data with 8:2 ratio. We determined the following hyper-parameters; batch size as 256, optimization function as "SGD" with a learning rate of 0.001 and momentum 0.9, and the number of epochs

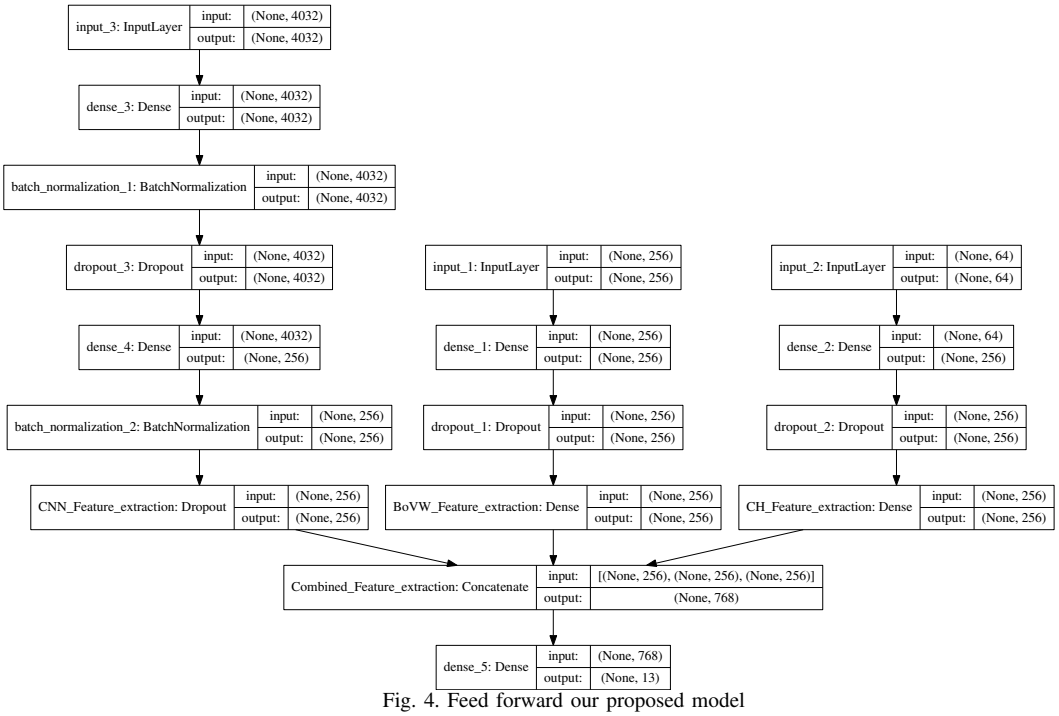

Fig. 4. Feed forward our proposed model

Algorithm 1 Predicting multi hot vector for an image

Input: Image data i including K kinds of sentiments
Output: Multi hot vector S_i

for k **in** range (K):
  for j **in** range (J):
    $Prob_{i,j,k}$=$FeatureExtraction_j(O_i(Z_k))$
  end for
end for
for j **in** range (J):
  T_i^k=mean$(\sum_j Prob_{i,j,k})$
  for j **in** range (J):
    $threshold_{j,k}$=median$(Prob_{i,j,k})$
  end for
  $All\_threshold_k$=mean$(\sum_j threshold_{j,k})$
  S_i^k=1 **if** $(T_i^k \geq All\_threshold_k)$ **else** S_i^k=0
end for

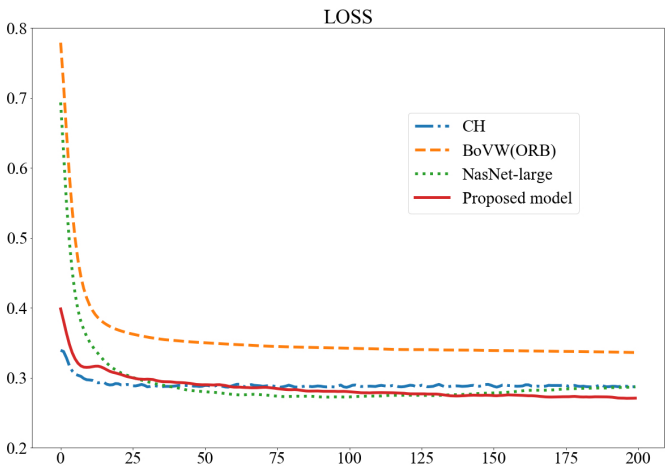

Fig. 5. Testing loss at four models (CH, BoVW, CNN, and Proposed model) in Transf13

200. For the implementation, we employ Keras [30] as our deep learning framework. Here, testing loss with four models (CH, BoVW, CNN, and our proposed model) in Transf13 is illustrated in Figure 5.

We compute the confusion matrix for Emoton6 and Transf13. We combine the most confusing classes into a single class and then observe the accuracies for the new classes. Here, the confusion matrix for Emotion6 and Transf13 models are both illustrated in Figure 6.

For the evaluations of multi-label classification, we employ two measures; exact match and mean average precision (mAP). Exact match also called subset accuracy indicates the percentage of samples that have all their labels classified correctly, while mean average precision for a given collection of unknown test data denotes the mean average precision scores. Table II shows the results. Here we compare Transf13's with Emotion6 both in terms of exact match and mean average precision. Also the table includes several base line methods including CH, BoVW (ORB), CH + BoVW (ORB), VGG16, ResNet50, and NasNet-large. The "Dimension" column of the table represents the feature dimension. For our proposed combined model, we have tested with three variations, i.e., CH+BoVW(ORB)+VGG16, CH+BoVW(ORB)+ResNet50, CH+BoVW(ORB)+NasNet-large. It turns out that

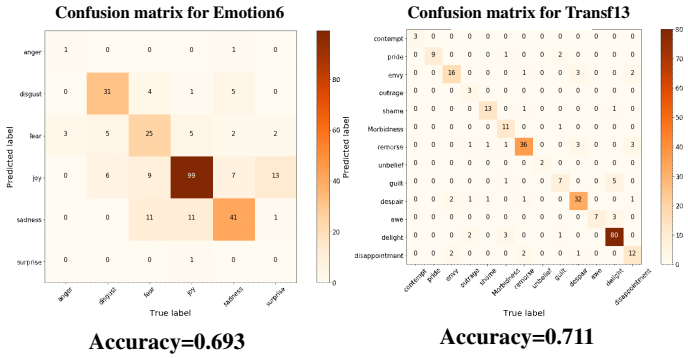

Fig. 6. Confusion matrix for Emoton6 and Transf13

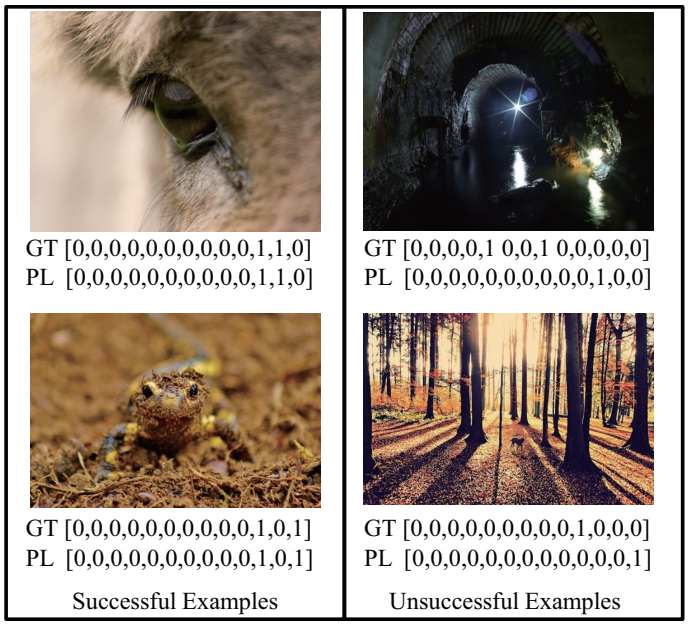

Fig. 7. Examples of experimental results
GT: Ground Truth, PL: Predicted Label.

CH+BoVW(ORB)+NasNet-large has the best mAP. Table III shows the results. Here we compare Transf13' s, Emotion6 and EmotionROI [31] both in terms of exact match and mean average precision.

With our proposed model, it is observed that by inflating the number of emotions from 6 to 13, both exact match and mAP scores of Transf13 have become better than those of Emotion6. As we predicted in Section III A, it seems that we could partially capture abundant psychological human feelings by infating emotions.

Figure 7 exhibits both successful and unsuccessful examples of our experiments. We could correctly predict multi-hot (multi-label) vectors for the left two pictures, while for the right two pictures we failed to predict the labels.

V. Conclusions

We proposed a model for visual sentiment analysis which accurately estimates multi-label multi-class problems from given images, evoking multiple different emotions simultaneously. We introduced the emotion inflation method from Emotion6 into 13 emotions (which we call 'Transf13') by means of emotional dyads. We performed multi-label sentiment analysis using the emotion-inflated dataset where we proposed a combined deep neural network model which enabled inputs to come from both hand-crafted features (e.g. BoVW (Bag of Visual Words) features) and CNN features. We also introduced a median-based multi-label prediction algorithm, where we assumed that each emotion had a probability distribution. Specifically, after training our deep neural network, we could predict the existence of an evoked emotion for a given unknown image if the intensity of the emotion was larger than the median of the corresponding emotion. Experimental results demonstrate that our Transf13-based model outperforms Emotion6-based model in terms of exact match (subset accuracy) and mean average precision. In addition, testing loss of our proposed model demonstrates that it outperforms other models. The mAP of our proposed model is higher than the mAP of the baseline model. We suggest that CH is affected by our proposed model. CH is one feature because CH is still a crucial factor in emotion recognition. There is a a close relationship between CH and sentiment [10]. In addition, saturation and brightness can have a direct impact on several sentiments.

Future directions might include the optimal weights for the linear combination of multiple neural networks given an arbitrary emotion evoked image dataset.

ACKNOWLEDGMENT

A part of this research was carried out with the support of the Grant-in-Aid for Scientific Research (B) (issue number 17H01746), and Grant for Education and Research in Toyohashi University of Technology.

References

[1] Yangyu Fan, Hansen Yang, Zuhe Li, and Shu Liu. Predicting image emotion distribution by emotional region. *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–9, 2018.

[2] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, Nov 2011.

[3] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 223–232, New York, NY, USA, 2013. ACM.

[4] Martin Solli and Reiner Lenz. Color based bags-of-emotions. In *International Conference on Computer Analysis of Images and Patterns*, pages 573–580. Springer, 2009.

[5] Marie Katsurai, Takahiro Ogawa, and Miki Haseyama. A cross-modal approach for extracting semantic relationships between concepts using tagged images. *IEEE Transactions on Multimedia*, 16(4):1059–1074, Jun 2014.

[6] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3 – 14, 2017. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

[7] Can Xu, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li. Visual sentiment prediction with deep convolutional neural networks. *ArXiv*, 11 2014.

TABLE II
THE RESULTS OF DOING EXPERIMENTS IN TRANSF13 AND EMOTION6 FOR MULTI-LABEL SENTIMENT CLASSIFICATION

		Emotion6		Transf13	
Method	Dimension	Exact match	mAP	Exact match	mAP
Base Line(BoVW(ORB)+NasNet-large)	512	0.152	0.407	0.520	0.627
Proposed Model (CH+BoVW(ORB)+NasNet-large)	768	0.146	0.581	0.574	**0.724**
CH	256	0.210	0.485	0.282	0.443
BoVW (ORB)	256	0.072	0.451	0.327	0.391
CH+BoVW(ORB)	512	0.035	0.416	0.039	0.304
VGG16	256	0.189	0.596	0.619	0.701
ResNet50	256	0.232	0.619	0.460	0.619
NasNet-large	256	0.197	0.611	0.520	0.646
CH+BoVW(ORB)+ResNet50	768	0.129	0.573	0.520	0.627
CH+BoVW(ORB)+VGG16	768	0.116	0.568	0.500	0.619

TABLE III
COMPARE WITH EMOTIONROI, TRANSF13 AND EMOTION6 OF MAP IN OUR PROPOSED METHOD.

Dataset	Exact match	mAP
Transf13	0.574	**0.724**
Emotion6	0.146	0.581
EmotionROI[31]	0.135	0.573

[8] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Macario O. Cordel, II, Shaojing Fan, Zhiqi Shen, and Mohan S. Kankanhalli. Emotion-aware human attention prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] H. Kim, Y. Kim, S. J. Kim, and I. Lee. Building emotional machines: Recognizing image emotions through deep neural networks. *IEEE Transactions on Multimedia*, 20(11):2980–2992, Nov 2018.

[11] Anan Liu, Yingdi Shi, Peiguang Jing, Jing Liu, and Yuting Su. Low-rank regularized multi-view inverse-covariance estimation for visual sentiment distribution prediction. *Journal of Visual Communication and Image Representation*, 57:243 – 252, 2018.

[12] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. Aligning where to see and what to tell: image caption with region-based attention and scene factorization, 2015.

[14] Atifa Athar, Muhammad Saleem Khan, Khalil Ahmed, Aiesha Ahmed, and Nida Suhail Anwar. A fuzzy inference system for synergy estimation of simultaneous emotion dynamics in agents. 2011.

[15] Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5413–5423. Curran Associates, Inc., 2017.

[16] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. 2008.

[17] K. Peng, T. Chen, A. Sadovnik, and A. Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, June 2015.

[18] Paul Ekman. What emotion categories or dimensions can observers judge from facial behavior? *Emotions in the human face*, pages 39–55, 1982.

[19] Robert Plutchik. *Emotion : theory, research, and experience*. Academic Press, 1980.

[20] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

[21] Agnese Augello, Ignazio Infantino, Giovanni Pilato, Riccardo Rizzo, and Filippo Vella. Binding representational spaces of colors and emotions for creativity. *Biologically Inspired Cognitive Architectures*, 5:64 – 71, 2013. Extended versions of selected papers from the Third Annual Meeting of the BICA Society (BICA 2012).

[22] G. CSURKA. Visual categorization with bags of keypoints. *Proc. of ECCV Workshop on Statistical Learning in Computer Vision, 2004*, pages 1–22, 2004.

[23] PM Panchal, SR Panchal, and SK Shah. A comparison of sift and surf. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2):323–327, 2013.

[24] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In Aleš Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, pages 430–443, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[25] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 778–792, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[29] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[30] François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[31] K. Peng, A. Sadovnik, A. Gallagher, and T. Chen. Where do emotions come from? predicting the emotion stimuli map. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 614–618, Sep. 2016.