Emotional Understanding of an Image by Applying High-level Concepts on Image Parts

Sharoon Nasim
Department of Computer Science
Forman Christian College
Lahore, Pakistan
sharoonnasim@fccollege.edu.pk

Mahnoor Rehan
Department of Computer Science
Forman Christian College
Lahore, Pakistan
213513461@formanite.fccollege.edu.pk

Nosheen Sabahat
Department of Computer Science
Forman Christian College
Lahore, Pakistan
nosheensabahat@fccollege.edu.pk

2020 IEEE 23rd International Multitopic Conference (INMIC) | 978-1-7281-9893-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/INMIC50486.2020.9318049

***Abstract*—This paper is the outcome of the research that has been conducted to investigate the relationship of applying High-Level Concepts (HLC) on a complete image and applying it along with its different parts. The implementation of technique includes the images that are divided into subparts followed by focusing on each part individually. To conclude our results, we combined the outcome of our mechanism and compared the results with the previous work. We have applied the object detection HLC on parts of an Image along with entire image. We also compared our results with previous findings and come up with the discovery of our results that have performed better than previously known work.**

***Index Terms*—High Level Concepts, VGG16, Emotion Classification**

I. Introduction

The interpretation of images is an emerging interest with remarkable innovations. In comparison to the complexity of the process of perceiving data by humans, the loophole for image understanding by the machines still exist. The advancements in the effective analysis of machines' emotional intelligence have taken the Human-Machine interaction to the next level [1]. However, the inaccuracy of results is due to the lack of focus on the combination, contract, and relativity between the contextual and visual features of the images [2].

Affective computing is an extremely advanced field. It recognizes human emotions, body language and facial expressions. Unlike the human cognitive approach, training machines like the human brain is a crucial part and still a challenge in emotional understanding. To do so, an effective application of affective analysis (algorithm) is required for better predictions of human observation and feedback for the viewed image [3]. Most of the Affect analyzers recognizes the emotional aspect of the individual subject only [1]. Multiple approaches are applied to identify the emotional states of each member in a group shown in an image. The combination of emotions of each group member is made through the Bottom-up level emotions and the Top-down group-level emotions are concluded from the social identification that affects the emotions of individual [1],[4],[5]. Facial expression recognition has achieved significant success but still, it is difficult to classify a complete image in any emotional class. To identify the human emotions, the total dependency, and performance is based on Low-level image feature (LLF) execution that is unsuccessful. The scene and object detection by using the High-Level Image feature (HLF) play a significant role in the classification of emotional states. In HLF, the set of features is used to detect the semantics (low-level features) and contextual information about the images. A hybrid approach is used, that is inspired by previous work [3], but the main difference is the technique and investigation of applying high-level concepts (pre-trained model) on image parts along with image and the complete image. The outcome has performed much better with a unique technique that is never been experimented and causes a sense of motivation in knowing the difference in performance with state of the art. The applied technique can be used to find better results using further HLCs in future works.

II. Related Work

"High-Level Concepts for Affective Understanding of Images" [3] is published in WACV 2017, they have used a Hybrid approach that has combined different High-level concepts (Object detection, Place detection) and low-level features for image understanding. They have applied the High-level concepts on the Emotion6 dataset and trained 7 regressors to foresee the distribution of each image and made predictions on 7 emotional categories (Surprise, Joy, Fear, Anger, Disgust, Sadness and Neutral). Fig. 1 shows the Means square error of these 7 regressors for each emotional distribution.

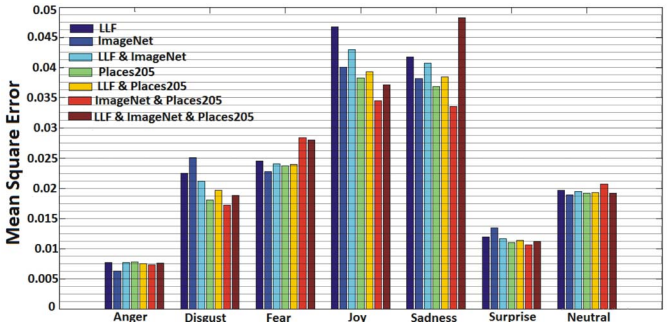

Fig. 1: Means square error of 7 regressor for each emotion class [3]

The regressors are chosen according to their efficiency to predict the probability distribution for all 7 emotional classes.

978-1-7281-9893-4/20/$31.00 ©2020 IEEE

The regressors that performed well for the respective classes were considered as a part of Hybrid Model. As shown in Fig.1, performances are inversely proportional to Mean Square Error. In this experiment, whichever regressor performs better to predict the probability distribution for any class, researchers have made that regressor part of the hybrid model.

Through their approach, they have got 52% accuracy for predicting correct classes for the complete dataset (1980 images). These predictions were chosen according to the dominance of the categorized class probabilities provided by a Hybrid model.

Faster R-CNN [6] is a well-known approach to detect the objects in real-time. This approach uses a VGG-16 pre-trained model and a Region Proposed Network. Though the idea of our approach is quite similar, our focus is not to detect the object instead we wanted to study the difference between applying HLC on complete image and specific parts of the image.

According to Kim et al. [7], the deep learning method is considered as an effective tool to recognize emotions. The combinations of objects and semantic background high-level features play a significant role in identifying the probability of multiple emotional states. Their approach was based on a feedforward deep neural network that produces the emotion values of a given image. The output emotion values in their framework are continuous in 2-dimensional space (valence and arousal. Valence values show the positivity or negativity of the image while arousal values show the excitement level of the image.

Peng et al. [8] assemble the Emotion6 database with multiple emotion labels, they introduced that one image has multiple emotions as Ekman described emotion classes [9]. The outcome of such predictions includes several responses to each category along with the most influential one. Though the combination of texture and color intensity explains the type of an image well in previous studies, this paper includes the types of emotional responses to such images. The images are modified to fit an image into a specified emotion category. Yang et al. [10] explains the intermediate approach that is applied to the attributes of Sentibank [11] and the Neural Networks [12] are trained with the conditional probability application. Unique algorithms were implemented to refine the probability distribution of emotions for each emotional state. Therefore, the Label Distribution Learning (LDL) method was selected that worked on the Conditional Probability Neural Network (CPNN) and was then trained. The results were impressive and were received from the 20,475 image datasets. Zhao et al. [13] predicted emotion distribution with the implementation of sparse coding. The objective of such a different approach is the same. To categorize the image types, some common features are selected, and sparse dictionary learning is enhanced for successful predictions. Such a top-quality approach was ahead than state-of-the-art methodologies.

Afsheen et al. [3] used Linear Hybrid Model that uses 7 regressors to predict emotional distribution probabilities following the emotional classes and studied the relationship between high-level concepts and emotions. All the methods above extract features from the whole image. Further improvement of 1980 images of Emotion6, Peng et al. developed an improved dataset EmotionROI [14], the ground truth is the Emotion Stimuli Map (ESM) which represents the intensity of emotions in accordance to the pixel densities. The level of emotion is recognized by the pixel accumulation intensities. The actual data was collected by the observers' major responses. After that, they used Fully Convolutional Networks with Euclidean Loss (FCNEL) to predict the ESM. Fan et al. [15] proposed a framework that uses ESM to crop images and used them to train SVR models and predict emotion distribution.

Fig. 2: Emotion Stimuli Map [14]

Bo et al. [16] presented their work for detecting Image Emotion by using a Deep Neural network. They extract the middle layer features (Itten contrast, figure-ground color difference, figure-ground area difference, and cool-color ratio) based on the art theory and combine them with the low-level features (saturation, light, dominant colors, texture) to predict image emotion. In concluding remarks, they suggested that High-level image features and a more suitable neural network model can be used to get better results.

In the proposed approach, we worked on the improvement of the categorical emotional predictions by applying the High-level concepts. Taking the previous work on the whole image into consideration, this unique approach of feature extraction on images include the partitioning of an image with one of the best performing object detection methods (VGG-16) [17]. The improvement in overall results is significant.

III. Dataset

The experiments reported in this paper use the well-known image datasets of Emotion6 [8], widely used for affective computing. This dataset is collected by the keyword of emotion class defined by Ekman [9]. Each class contains 330 images. After collecting 1980 images (330 x 6), they performed a user study through which they obtained probability distribution vectors for each image. Every image has a 7-dimensional emotion probability distribution vector (for six Ekman's emotion classes and an additional neutral class) [9]. Valence arousal (VA) values are also part of each vector. Each image is no longer associated with a single emotion class. Even in reality, a person has multiple emotions toward one image [18].
For classification experiments, we consider the emotion class which has the highest probability in the emotion distribution provided as ground truth for each image. Table I shows Image classification based on the highest probability distribution that has been provided in the ground truth.

TABLE I: Image Classification

Dataset	*Emotion6 (1980 Images)*
Anger	31
Fear	329
Disgust	245
Joy	638
Sadness	308
Surprise	104
Neutral	325

IV. Methodology

Our experimentation work is divided into several modules. Fig.3 shows the flow diagram of our proposed methodology.

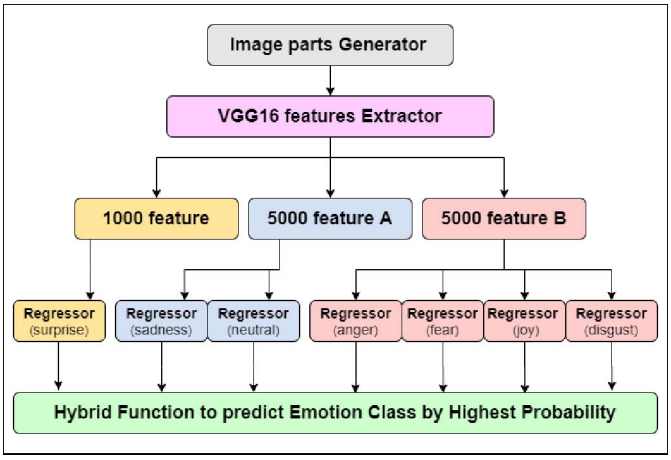

Fig. 3: Flow Diagram that shows the experimentation framework

A. Image Part Generator

In this module, we processed each of our images. The first step is to resize all images into 400 x 400 pixels. Later, we cropped each image into four sub-images sized 200 x 200 pixels. After having a complete image along with those 4 parts (shown in Fig.4). The two schemes for the image sub-part division are explained in the following way:

- Scheme A: We have cropped top-middle, bottom-middle, middle-left, middle-right.
- Scheme B: We divided an image into 2 x 2 grid, making its 4 sub-parts: Top-left, Top-right, Bottom-right, Bottom-left.

Now we have five image arrays in each scheme having four sub-parts along with a complete image. For understanding an image, we focus on its object. By applying HLC on a single image, it focuses on a single object. Our approach highlights at least four to five objects.

B. High Level Concept

As HLC, we have used VGG-16 [17] the object detection model, that is trained over 14 million images from 1000 categories. This model works with minimum input of an image containing 224 x 224 pixels and a return array of 1000 features. The images are resized into 224 x 224 pixels followed by the application of HLC's on all 9 images generating three feature sets.

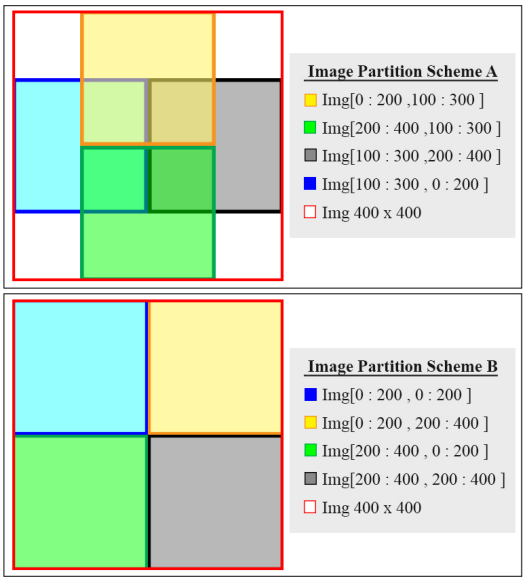

Fig. 4: Focus Parts of the Image

C. Features Set

- Complete Image: By applying HLC on a complete image, we have got an array of 1000 feature.
- Feature Set A: By combining the HLC features of the scheme A and complete image, we have an array of 5000 features.
- Feature Set B: By combining the HLC features of scheme B and the complete image, we have an array of 5000 features.

D. Hybrid Model

We have trained seven ridge regressors overall feature sets to predict the probability distribution for each emotion class. To associate an image to an emotional class, we have created a hybrid model that predicts the probability distribution for all seven emotional classes by opting for the best predicting regressors. Based on the probability distribution, it is easy to classify an image into an emotional category by considering the probability distribution of the highest dominating class.

V. Results & Analysis

In experimentation, the performance of regressors is observed for overall feature sets. To analyze the error of prediction and ground truth probabilities, we have used the mean square error as a performance measure. Table II shows the results for these regressors. The less mean square error shows better results.

TABLE II: Performance for Probability Distribution.

Distribution	1000 feature		5000 feature A		5000 feature B	
	MSE	α	MSE	α	MSE	α
Anger	0.0053	1	0.0053	10	**0.00524**	10
Fear	0.0248	1	0.0245	1	**0.0244**	1
Surprise	**0.01224**	1	0.01226	10	0.01226	1
Joy	0.0346	1	0.0319	1	**0.03135**	1
Disgust	0.0200	1	0.0189	1	**0.0184**	1
Sadness	0.0323	1	**0.0306**	1	0.0311	1
Neutral	0.0183	10	**0.01707**	1	0.0178	10

Furthermore, we also predicted the valence and arousal points for each image taking mean absolute error as a performance measure. The purpose of calculating the mean absolute error is to compare our results with the previous work. Table III. shows the mean absolute error for predicted valence and arousal points against all feature sets.

TABLE III: Performance over VA-Score.

Categories	1000 feature		5000 feature A		5000 feature B	
	MAE	α	MAE	α	MAE	α
Valence	1.2408	0.1	1.1739	1	**1.1489**	1
Arousal	0.6839	1	0.6805	10	**0.6778**	10

VI. Comparison with Previous Work

The classification accuracy is measured by comparing the most dominant emotional probabilities from the ground truth and our results. Our Hybrid Model has achieved 66.7% accuracy in the classification task. Table IV shows the comparison of our Hybrid model with the previous work [3][15]. Moreover, we compared our model in terms of Kullback-Leibler divergence (KL) & Bhattacharya coefficient (BC) and normalized our probability distribution data in the following manner.

- Firstly, the negative probability distribution is replaced with zero.
- Secondly, zero's of the data are converted into 10^{-10} to avoid zero division error.
- Lastly, the scaling of predicted 7-dimensional output by our hybrid model adjusts it in such a way that their sum equals 1.

Table IV displays summary of the results. It presents the overall accuracy of the classification task, and also Kullback-Leibler divergence (KL) and Bhattacharya coefficient (BC) for the prediction of the probability distribution of our Hybrid model. The small KL value and the large BC value represents improved performance respectively.

TABLE IV: Summary of our Results and previous Results [3][15]

Approach	Accuracy (%)	KL	BC
ImageNet[3]	45.83	0.559	0.818
Hybrid Model[3]	52.00	0.493	0.839
With ESM [15]	-	0.480	0.851
Our Approach	**66.71**	**0.311**	**0.883**

We compared our hybrid model's regressors with the previous work [3]. Fig. 5 shows the comparison and presents the previous work in Fig. 1. This comparison depicts the difference between our approach and the improved performance of (VGG-16) HLC. Less mean square error (MSE) indicates better performance.

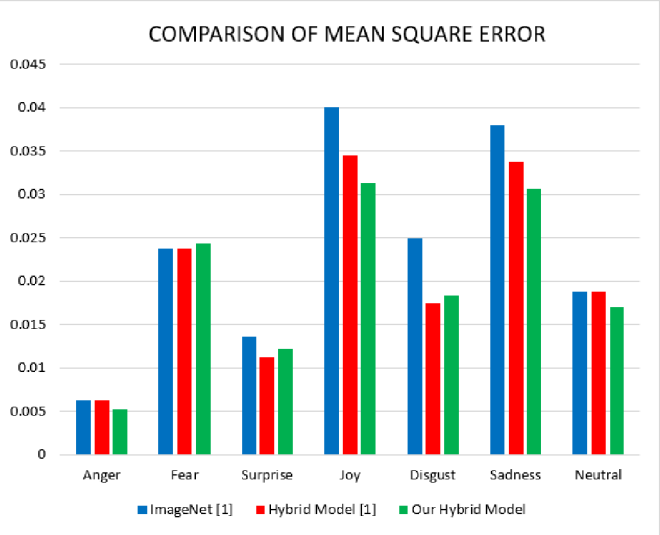

Fig. 5: Comparison of Mean Square Error

Table V summarizes the comparison of predicting valence arousal values with previous work [3], Mean Absolute error (MAE) is used to show the improved performance of our work. Less MAE value presents the efficiency of the approach.

TABLE V: Comparison of Mean Absolute Error between the previous result and the best results obtained by our approach.

Categories	Previous Results (MAE)[3]		Our Approach	
	ImageNet	Best	MAE	α
Valence	1.5851	1.2093	1.1489	1
Arousal	0.6898	0.6802	0.6778	10

VII. Conclusion

The paper is consolidated with the implementation of High-level concepts for feature extraction and the performance is observed. It involves the object detection (VGG-16) strategy differences on a complete image and parts of an image. The experimentation reports show better performance for feature set A and B. This technique works differently for each emotion class. Multiple tables present the summary of our findings with multiple aspects. Most of our work is inspired by "High-Level Concepts for Affective Understanding of Images" [3]. Therefore, we compared our results with the results from [3][15].

Our Hybrid Model efficiently predicts the probabilities for each emotion class. By finding the dominant emotional response in the 7-dimensional output of our hybrid model, this approach has scored 66.7% accuracy in classification which is better than the previous result [3] reported on the Emotion6 dataset [8]. Additionally, we also performed Valence Arousal scores comparisons over the Emotion6 dataset, shown in Table V.

In future, we intend to substantially explore the diversification of HLCs taking time detection, and places detection into consideration. In contrast to the VGG-16 Model, the success rate and efficacy of such an upgraded model is considerable. Further improvement can also be done on the image part generator. The refined approach of object detection on image partition is compared with object detection on a single and complete one. With the improved results, we envision the further experimentation of techniques on HLC's.

References

[1] Wenxuan Mou, O. Celiktutan and H. Gunes, "Group-level arousal and valence recognition in static images: Face, body and context," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, 2015, pp. 1-6, doi: 10.1109/FG.2015.7284862.

[2] M. Xua, C. Xub, X. Hea, J. Jinc, S. Luoc, and Y. Ruid, "Hierarchical Affective Content Analysis In Arousal And Valence Dimensions," Signal Processing, vol. 93, no. 8, pp. 2140–2150, 2013.

[3] A. R. Ali, U. Shahid, M. Ali and J. Ho, "High-Level Concepts for Affective Understanding of Images," 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, 2017, pp. 679-687.

[4] S. G. Bars¨ade and D. E. Gibson, "Group emotion: A view from top and bottom," Research on managing groups and teams, vol. 1, pp.81–102, 1998.

[5] E. R. Smith, C. R. Seger, and D. M. Mackie, "Can emotions be truly group level? evidence regarding four conceptual criteria." Journal of personality and social psychology, vol. 93, no. 3, p. 431, 2007.

[6] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

[7] Kim, H.-R., Kim, Y.-S., Kim, S. J., & Lee, I.-K. (2018). Building Emotional Machines: Recognizing Image Emotions through Deep Neural Networks. IEEE Transactions on Multimedia, 1–1. doi:10.1109/tmm.2018.2827782

[8] K.-C. 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. IEEE, jun 2015.

[9] P. Ekman. An argument for basic emotions. Cognition & Emotion, 6(3-4):169–200, May 1992.

[10] J. Yang, M. Sun, and X. Sun, "Learning visual sentiment distributions via augmented conditional probability neural network," in AAAI Conference on Artificial Intelligence, 2017, pp. 224–230 .

[11] Borth, et al. "Large-scale visual sentiment ontology and detectors using adjective noun pairs." 2013, pp. 223-232.

[12] Geng, X., C. Yin, and Z. H. Zhou. "Facial Age Estimation by Learning from Label Distributions," IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, pp. 2401-2412.

[13] Zhao, Sicheng, et al. "Predicting discrete probability distribution of image emotions," IEEE International Conference on Image Processing IEEE, 2015, pp. 2459-2463.

[14] Peng, Kuan Chuan, et al. "Where do emotions come from? Predicting the Emotion Stimuli Map," IEEE International Conference on Image Processing IEEE, 2016, pp. 614-618.

[15] Fan, Y., Yang, H., Li, Z., & Liu, S. (2018). Predicting Image Emotion Distribution by Emotional Region. 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI).

[16] Li, B., Guo, C., & Ren, H. (2018). Image Emotion Recognition Based on Deep Neural Network. 2018 IEEE International Conference of Safety Produce Informatization (IICSPI). doi:10.1109/iicspi.2018.8690404

[17] Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

[18] Y. Fan, H. Yang, Z. Li and S. Liu, "Predicting Image Emotion Distribution by Learning Labels' Correlation," in IEEE Access, vol. 7, pp. 129997-130007, 2019, doi: 10.1109/ACCESS.2019.2939681.