Fuzzy Systems and Data Mining VII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210176

Facial Expression Recognition and Image Description Generation in Vietnamese

Khang Nhut LAM [a,1], Kim-Ngoc Thi NGUYEN [a], Loc Huu NGUY [a], and Jugal KALITA [b]

[a] *Can Tho University, Can Tho, Vietnam*
[b] *University of Colorado, Colorado Springs, USA*

Abstract. This paper discusses a facial expression recognition model and a description generation model to build descriptive sentences for images and facial expressions of people in images. Our study shows that YOLOv5 achieves better results than a traditional CNN for all emotions on the KDEF dataset. In particular, the accuracies of the CNN and YOLOv5 models for emotion recognition are 0.853 and 0.938, respectively. A model for generating descriptions for images based on a merged architecture is proposed using VGG16 with the descriptions encoded over an LSTM model. YOLOv5 is also used to recognize dominant colors of objects in the images and correct the color words in the descriptions generated if it is necessary. If the description contains words referring to a person, we recognize the emotion of the person in the image. Finally, we combine the results of all models to create sentences that describe the visual content and the human emotions in the images. Experimental results on the Flickr8k dataset in Vietnamese achieve BLEU-1, BLEU-2, BLEU-3, BLEU-4 scores of 0.628; 0.425; 0.280; and 0.174, respectively.

Keywords. facial expression recognition, image description, CNN, YOLOv5, VGG16, LSTM.

1. Introduction

Facial expression recognition (FER) and image description generation (IDG) are important tasks in image understanding, connecting computer vision with natural language processing. Image content can be partially described through objects and their locations. Huang et al. [1] classify approaches for FER into 2 groups, including conventional FER and deep learning based approaches. Given an input image, the conventional FER approach consists of several steps: pre-processing the image to reduce noise (e.g., Gaussian Filter [2], Bilateral Filter [3]), detecting face and facial components, extracting features (e.g., Local Directional Pattern [4], Histogram of Oriented Gradients [5]), and classifying emotions (e.g., Support Vector Machine [6], Naïve Bayes classifier [7]). Li and Deng [8] report that deep FER consists of several steps including face alignment detectors (e.g., Viola-Jones [9], face alignment 3000fps [10]), data augmentation (e.g., rotation, skew, scaling), face normalization, feature learning, and emotion classification. Several deep neural network models have been used to learn image features such as Convolutional Neural Network (CNN) [11,12], the hybrid Convolution - RNN [13], and Generative Ad-

[1]Corresponding Author: Khang Nhut Lam, Can Tho University, Campus II, 3/2 Street, Can Tho City, Vietnam; E-mail: lnkhang@ctu.edu.vn.

versarial Network [14]. Human facial expressions are usually classified into 7 categories [15]: afraid, angry, disgusted, happy, neutral, sad, and surprised.

Publishing efforts on IDG may be grouped into 3 categories [16,17]: (i) Models rely on computer vision techniques to identify objects in the input image and extract their features [18,19]. Features extracted are fed to a Natural Language Generation (NLG) [20] sub-system. Then, the steps to build a description for the image follow the NLG architecture. (ii) Models are based on a retrieval system, where the image descriptor is retrieved from the training dataset. Most of these systems use neural models to extract image features and linguistic information [21,22]. (iii) A system relies on the generation architecture to generate new descriptions. First, neural models are used to extract features of images (e.g., VGGNet [23], Faster R-CNN [24], ResNet [25], and Inception-V3 [26]), then neural models are used to generate new descriptions [27,28]. For the last category, there are 2 architectures to generate images captions: inject and merge architectures [29]. In the inject architecture, the vectors of image features and words are combined and fed into a neural model for generating image captions; whereas in the merge architecture, the image feature vectors are merged with the final state of the neural model in a multimodal layer. The experiments show that the merge architecture outperforms the inject architecture.

The more information provided in the descriptive sentences, the image caption model is better. For example, the image caption "A boy with a happy face in a red shirt is playing on grass" is more detailed and vivid than the image caption "'A boy is playing on the grass". We have not seen image captioning with emotion recognition. This paper aims to explore the methods for FER and IDG. If the sentence describing the image contains words or phrases referring to a human, the system will identify the facial expression and add this emotion to the image description. In other words, the description sentence describes the content in the image and the facial emotion of the person.

2. Proposed Approach

We first discuss the datasets used and methods to pre-process datasets. Then, we present approaches to recognize human facial expressions and generate descriptions of images.

2.1. Datasets Pre-processing

The KDEF [15], Flickr8k [30], and Flickr30k[31] datasets are used to train the models to recognize the human facial expressions and to generate the image descriptions, respectively. The KDEF dataset comprises 4,900 images of 70 individuals displaying 7 emotional expressions, each of which is viewed from 5 different angles. The dataset is divided by 70%, 20%, and 10% for training, validation, and test sets, respectively. The Flickr8k and Flickr30k datasets consist of 8,092 images and 31,783 images, respectively, each of which has 5 description sentences. Each dataset is divided into 2 parts: 1,000 images for testing, the rest of the images for training.

The description sentences in the Flickr8k dataset are in English. We translate this dataset to Vietnamese using a pre-trained Transformer model[2]. The Transformer translation model was trained on the dataset using 600,000 sentences extracted from TED[3]. Then, we pre-process the description sentences by converting sentences to lowercases, removing special characters and punctuation marks. The Underthesea[4] toolkit is used

[2]https://github.com/pbcquoc/transformer
[3]https://www.ted.com/
[4]https://pypi.org/project/underthesea/

to segment words in the description sentences. The dictionary created consists of 4,028 words. Image descriptions are embedded in terms of vectors based on the position of words in the dictionary.

2.2. CNN-based and YOLOv5-based FER Models

The CNN model [32] has 3 types of layers, including convolution, pooling, and fully connected layers. The convolution layer extracts feature maps from input images by using filters to perform convolution operations. Then, the features extracted are applied nonlinear transfer functions such as ReLU, sigmoid, tanh, and softmax. The pooling layer reduces the dimensionality of the output of the previous layer by performing a pooling operation such as max pooling, average pooling, and sum pooling. Finally, the fully connected layer or dense layer is a normal flat feed-forward neural network layer using a nonlinear activation function to obtain the probability of each class. In this paper, we use a classic feed-forward CNN to detect facial expressions. The OpenCV library is used to detect and crop human faces in each image, and then convert face images to grayscale. The grayscale images are fed to the CNN model using ReLU activation function, average pooling operation for training FER.

A traditional CNN does not detect and label objects well in real-time. YOLO [33] is state-of-the-art in real-time detecting objects. YOLOv3 can predict the bounding box and process the image simultaneously, so it is less time-consuming. The accuracies of using YOLOv3 for facial expression recognition on JAFFE, RaFD, and CK+ are 98.12%, 97.01%, and 99.72% [34], respectively. Currently, the newest version YOLOv5 comprises 3 main components: Cross Stage Partial Network [35] (CSP) backbone (including CSPRestNext50 and CSPDarknet) for feature extraction, PA-NET [36] neck for feature aggregation, and Head with YOLO layer for predicting boxes and labels. In our FER experiment, we use YOLOv5 provided by Ultralytics[5].

2.3. Image Description Generation Model

We use the merge architecture to construct descriptions for images [29,37] with the VGG16 model for extracting image features and the LSTM model for constructing image caption, as presented in Figure 1. In our implementation, we use the library toolkits supported by TensorFlow.

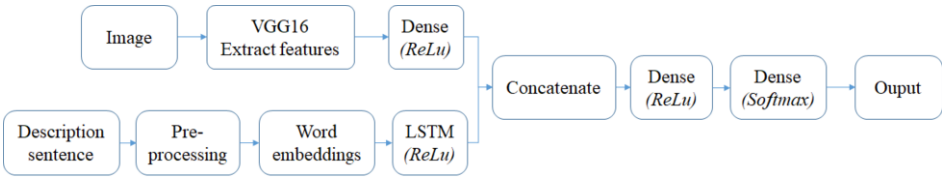

Figure 1. The image description generation model [29,37]

The VGG16 model extracts image feature vectors of size 4,096, which then are fed into a dense layer using the ReLU activation function. Description sentences are preprocessed, segmented, padded, and embedded into vectors. The word embeddings are fed to the LSTM using ReLU function. The outputs of the dense neural network and the LSTM have a similar size of 256 and then are concatenated and passed through a Dense layer using ReLU function with an output space of 256 and the next Dense layer using the softmax activation function with an output space of the dictionary size of 4,028.

[5]https://ultralytics.com/yolov5

If there is a personal noun in the caption generated, we simply pass the human image into the FER model and add the result of the model to the description generated. A list of person nouns is constructed manually including {đứa trẻ (kid), người đàn ông (man, male), cậu bé (boy), thanh niên (adolescent), trẻ em (baby), cô bé (girl), người phụ nữ (woman, female, lady), chàng trai (boy), ông già (old man), bà già (old woman), em bé (baby), bé gái (girl)}.

During the experiment, we noticed that sometimes the model did not perform correctly in determining the color of the object in the image. To solve the problem of color recognition (CR), from the built-in descriptive sentence, if the sentence contains color words, we determine the name of the object that needs color recognition. We use the YOLOv5 model to locate the object, crop the object, and save it as a new image for color recognition. An object usually has more than one color, we extract the dominant color of the object following the instructions of Ercolanelli[6]. When cropping an object out of an image, the object is usually in the center of the image, so the colors in the four corners of the image are usually the background colors. Therefore, pixels close to the color of the four corners of the image are considered the background color and excluded from the image. Then, a clustering algorithm, the K-means algorithm, is used to group similar pixels. The most dominant color of the object is considered the color of that object. Finally, the K Nearest Neighbors algorithm is used to convert the object color to words in human language by finding the nearest neighbor color in a large dictionary of colors taken from the XKCD color survey. The color after identification is replaced with the color of the previous description. Figure 2 shows an example of generating an image description.

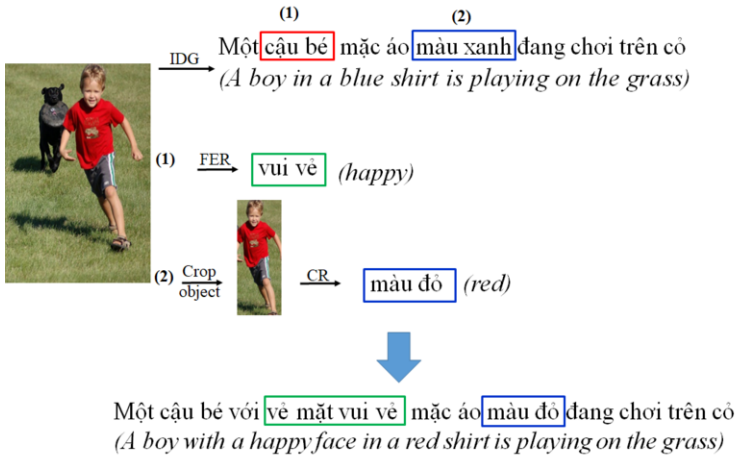

Figure 2. Example of generating an image description

3. Experimental Results

We build the modes in the Google Colab environment with 12GB RAM with GPU. The FER test set has 490 images with 78 afraid (meaning "sợ hãi"), 83 angry (meaning "giận dữ"), 71 disgusted (meaning "ghê tởm"), 65 happy (meaning "vui vẻ "), 66 neutral (meaning "trung lập"), 64 sad (meaning "buồn bã"), and 63 surprised (meaning "ngạc nhiên") emotions. The results of the FER models are presented in Table 1.

[6]https://github.com/algolia/color-extractor

Table 1. Recall, precision and F1-score of the FER models

Facial emotions	Recall		Precision		F1-score	
	CNN	YOLOv5	CNN	YOLOv5	CNN	YOLOv5
Afraid	0.705	0.897	0.785	0.921	0.742	0.908
Angry	0.795	0.939	0.942	0.975	0.862	0.956
Disgusted	0.859	0.901	0.813	0.941	0.835	0.920
Happy	0.953	0.969	0.968	0.984	0.960	0.976
Neutral	0.954	0.984	0.851	0.984	0.899	0.984
Sad	0.812	0.937	0.776	0.845	0.793	0.888
Surprised	0.920	0.952	0.828	0.923	0.871	0.937

The accuracies of the CNN and YOLOv5 models are 0.853 and 0.938, respectively. YOLOv5 recognizes human facial emotions better and faster than CNN in all emotions. The happy emotion can be recognized very well; whereas, the afraid emotion might be mis-recognized as the surprise emotion by both two emotion recognition models. Next, we evaluate the method for generating image descriptions in Vietnamese. Table 2 presents the BLEU-scores of the image description generation model (the so-called IDG) and the image description generation model with facial expression recognition (the so-called IDG with FER). Interestingly, the results show that the Flickr30k dataset, including more images and captions than the Flickr8k dataset, does not help achieve better BLEU scores. Some examples of the image descriptions are shown in Table 3.

Table 2. BLEU-scores of the image description generation models

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Flickr8k	IDG	0.629	0.426	0.281	0.175
	IDG with FER	0.628	0.425	0.280	0.174
Flickr30k	IDG	0.616	0.396	0.242	0.136
	IDG with FER	0.615	0.396	0.241	0.135

4. Conclusion

We experiment with CNN and YOLOv5 to recognize facial expressions on the KDEF dataset. The image description generation model integrated with the emotion recognition model and color recognition model achieves acceptable BLEU scores on the Flickr8k dataset. The description sentences generated by the current model can describe one person in an image. For future work, we will study approaches that might describe many people and their emotions in the image. Currently, we are performing experiments using Inception-V3 and YOLOv5 to extract image features instead of the VGG16 model, and training the models on different datasets. Besides, the Transformer model [38] or BERT model [39] will be used to generate description sentences. In addition, we need to improve the quality of the training dataset by improving the translation model.

References

[1] Huang Y, Chen F, Lv S, Wang X. Facial expression recognition: A survey. Symmetry. 2019 Oct; 11(10):1189.

[2] Deng G, Cahill LW. An adaptive Gaussian filter for noise reduction and edge detection. In IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference; 1993. p. 1615-1619.

Table 3. Examples of the image descriptions

Image	Image description	Image description with correction
	Một người đàn ông đang đi xe đạp trên một con đường (meaning "A man is riding a bicycle on a road").	None
	Một con chó đen và trắng đang chạy qua một cánh đồng (meaning "A black and white dog is running through a field").	Một con chó nâu đang chạy qua một cánh đồng (meaning "A brown dog is running through a field").
	Một cậu bé đang chơi trong hồ bơi (meaning "A boy is playing in the pool").	Một cậu bé với vẻ mặt vui vẻ đang chơi trong hồ bơi (meaning "A boy with a happy facial expression is playing in the pool").

[3] Zhang M. Bilateral filter in image processing. Master's Thesis, Louisiana State University, Baton Rouge, LA; 2009.

[4] Jabid T, Kabir MH, Chae O. Facial expression recognition using local directional pattern (LDP). In IEEE International Conference on Image Processing; 2010. p. 1605-1608.

[5] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 2005; Vol. 1, p. 886-893.

[6] Eng SK, Ali H, Cheah AY, Chong YF. Facial expression recognition in JAFFE and KDEF Datasets using histogram of oriented gradients and support vector machine. In IOP Conference series: materials science and engineering; IOP Publishing; 2019; Vol. 705, No. 1. p. 012031.

[7] Mao Q, Rao Q, Yu Y, Dong M. Hierarchical Bayesian theme models for multipose facial expression recognition. IEEE Transactions on Multimedia. 2016; 19(4). p.861-873.

[8] Li S, Deng W. Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing. 2020.

[9] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2001; Vol. 1.

[10] Ren S, Cao X, Wei Y, Sun J. Face alignment at 3000 fps via regressing local binary features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 1685-1692.

[11] Walecki R, Rudovic O, Pavlovic V, Schuller B, Pantic M. Deep structured learning for facial expression intensity estimation. Image Vis. Comput, 259; 2017. p.143-154.

[12] Pranav E, Kamal S, Chandran CS, Supriya M.H. Facial emotion recognition using Deep Convolutional Neural Network. In Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS); 2020. p. 317-320.

[13] Jain N, Kumar S, Kumar A, Shamsolmoali P, Zareapoor M. Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters; 2018; 115. p.101-106.

[14] Yang H, Zhu K, Huang D, Li H, Wang Y, Chen L. Intensity enhancement via GAN for multimodal face expression recognition. Neurocomputing; 2021; 454. p.124-134.

[15] Lundqvist D, Flykt A, Öhman A. The Karolinska directed emotional faces (KDEF). CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet; 1998; 91(630).

[16] Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler-Cinbis N, Keller F, Muscat A, Plank B. Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research. 2016; 55. p.409-442.

[17] Hossain MZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR). 2019; 51(6). p.1-36.
[18] Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg AC, Berg TL. Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence; 2016; 35(12). p. 2891-2903.
[19] Elliott D, Keller F. Image description using visual dependency representations. In Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2013. p. 1292-1302.
[20] Reiter E, Dale R. Building applied natural language generation systems. Natural Language Engineering. 1997; 3(1). p.57-87.
[21] Socher R, Karpathy A, Le QV, Manning CD, Ng AY. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics. 2014; 2. p.207-218.
[22] Sur C. Gaussian smoothen semantic features (GSSF)–exploring the linguistic aspects of visual captioning in Indian languages (Bengali) using MSCOCO framework. arXiv preprint arXiv:2002.06701. 2020.
[23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.
[24] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. 2015; 28. p.91-99.
[25] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-778.
[26] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 2818-2826.
[27] Aneja J, Deshpande A, Schwing AG. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 5561-5570.
[28] Guo L, Liu J, Zhu X, Yao P, Lu S, Lu H. Normalized and geometry-aware self-attention network for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 10327-10336.
[29] Tanti M, Gatt A, Camilleri KP. What is the role of recurrent neural networks (RNNs) in an image caption generator? arXiv preprint arXiv:1708.02043. 2017
[30] Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research. 2013; 47. p.853-899.
[31] Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics. 2014; 2. p.67-78.
[32] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. In Proceedings of the IEEE; 1998; 86(11). p.2278-2324.
[33] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 779-788.
[34] Luh GC, Wu HB, Yong YT, Lai YJ, Chen YH. Facial expression based emotion recognition employing YOLOv3 deep neural networks. In 2019 International Conference on Machine Learning and Cybernetics (ICMLC); 2019. p. 1-7.
[35] Wang CY, Liao HYM, Wu YH, Chen PY, Hsieh JW, Yeh IH. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020. p. 390-391.
[36] Wang K, Liew JH, Zou Y, Zhou D, Feng J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 9197-9206.
[37] Sharma G, Kalena P, Malde N, Nair A, Parkar S. Visual image caption generator using deep learning. In Proceeding of the 2nd International Conference on Advances in Science & Technology (ICAST); 2019.
[38] Wolf T, Chaumond J, Debut L, Sanh V, Delangue C, Moi A, Cistac P, Funtowicz M, Davison J, Shleifer S, Louf R. Transformers: State-of-the-art natural language processing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 2020. p. 38-45.
[39] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018