Continuous Learning Based Novelty Aware Emotion Recognition System

Mijanur Palash, Bharat Bhargava

Department of Computer Science, Purdue University, USA
{mpalash, bbshail}@purdue.edu

1 Abstract

Current works in human emotion recognition follow the traditional closed learning approach governed by rigid rules without any consideration of novelty. Classification models are trained on some collected datasets and expected to have the same data distribution in the real-world deployment. Due to the fluid and constantly changing nature of the world we live in, it is possible to have unexpected and novel sample distribution which can lead the model to fail. Hence, in this work, we propose a continuous learning based approach to deal with novelty in the automatic emotion recognition task.

2 Introduction

Understanding emotions properly is an important requirement for meaningful human communication. In addition, emotional state affects things like how a person drives a car, how a kid learns in the classroom or how police interact with a criminal on the road etc. Automatic emotion recognition has widespread usage in different areas such as human-computer interactions, law enforcement and surveillance, interactive gaming, consumer behaviour analysis, customer service, education, and health care etc.

Researchers in psychology categorized basic human emotions into anger, happiness, sadness, disgust, fear, contempt and surprise (Patel et al. 2020) classes. Human expresses these emotions using various verbal and non-verbal cues. Facial expression analysis is one of the major ways to identify the underlying emotion. Apart from that, various important conclusions for emotion can be drawn from posture and gait. Similarly, speech, writing, brain scan, EEG signal etc. also convey useful emotional information.

The advancement of machine learning and deep learning enables researchers to design more accurate classifiers for automatic emotion recognition systems. Researchers nowadays frequently use various ML and DL techniques to detect and classify emotions. Popular models include variations of support vector machine (SVM), deep belief network (DBN), ensemble learning, deep neural network (DNN) and convolutional neural network (CNN) etc. These models are trained on various datasets collected from web scrapping or created using volunteers and involve somebody to manually label the data samples. More details about some of the datasets are discussed in section 4.

However, this closed learning approach faces challenges in real-world situations where an unexpected sample may appear at any moment. A perfect machine learning model requires a perfect training dataset that truly represents all possible real-world situations. However, creating an all-inclusive dataset like this is very hard. Moreover, an emotion detection model trained on acted datasets where the margin between true and false samples are wide and clear may have a hard time correctly identifying emotion in a situation where someone is intentionally obfuscating his emotional state and hence the margin is thin and blurry. Another type of challenge arises when the model encounters a different type of emotion class absent in the training dataset as most of the datasets do not cover all of the emotion classes. In this work we define these types of situations as a novelty.

However, there is a lack of proper focus on handling novelty in current emotion recognition research. Researchers are more interested in coming up with newer models and newer datasets to get higher accuracy on the test data. However, we also need to look for answers to the questions such as how to detect a novelty, how to characterize it and how to adapt the systems to handle the novelties etc.

The main contributions of this work are:

1. We present the novelty aspects of the automatic emotion recognition task.
2. We present a system that addresses novelty in a continuous long learning manner for emotion recognition.
3. We present an emotion recognition classifier that detects emotion from facial image
4. We present a novelty detector that works in parallel with the classifier to detect novelty
5. We present experimental results and discuss potential next steps for this work.

3 Background and Related Work

Convolutional Neural Network (CNN) (LeCun et al. 1999) is a deep learning architecture widely used in image analysis. A CNN typically consists of convolutional, pooling and fully connected layers stacked together. The convolution can be defined as a repeated filter operation to an input that results in a feature map. For image classification, we can say CNN highlights important parts of the image to differentiate the classes.

Jadhav et. al. (Jadhav and Ghadekar 2018) used a convolutional neural network (CNN) to detect emotion from the facial expressions on the FER-2013 dataset. They achieved only 63% accuracy on the test dataset. This lower accuracy can be attributed to the simplicity of their model.

Gan et. al. (Gan, Chen, and Xu 2019) achieved improved accuracy on the FER-2013 dataset using ensemble CNN and a novel label perturbation strategy. Due to the enhanced discrimination ability of the ensemble method, they received higher accuracy of 73.7%.

Dhankar et. al. (Dhankhar 2019) used the ensemble method and transfer learning with VGG16 and RESENT-50 to overcome the limitations of the basic CNN method. However, their work did not achieve the best result and only scored 67% accuracy on the FER-2013. This may be attributed to a sub-optimal hyperparameter tuning.

Besides CNN, the support vector machine (SVM) is also popular in facial emotion recognition due to its lightweight architecture compared to CNN. Support Vector Machine (SVM) is a type of machine learning model which tries to identify the maximum margin plane between the classes. Datta et. al. (Datta, Sen, and Balasubramanian 2017) used SVM for emotion classification. Kurup et. al. (Kurup, Ajith, and Ramón 2019) used the deep belief network (DBN) technique for emotion classification.

4 Datasets

In literature, several datasets have been proposed for emotion recognition. A summary of notable datasets is presented in table 1.

Table 1: Different datasets for automatic emotion recognition. Here, N: Neutral, S: Sadness, Sr: Surprise, H: Happiness, F: Fear, A: Anger, B: Boredom, P: Puzzlement, Ax: Anxiety, C: Contempt and D: Disgust.

Name	# Items	Type	Setting	Classes
CK+	593	Video	Posed & spontaneous	N, S, Sr, H, F, A and D
FER-2013	32,298	Image	Posed	N, S, Sr, H, F, A and D
Emotic	23,571	Image	Wild	N, S, Sr, H, F, A, D and 19 other classes
AffectNets	450,000	Image	Wild	N, S, Sr, H, F, A, D and C
CAER-S	70,000	Image	TV shows	N, S, Sr, H, F, A and D
FABO	206	Video	Posed	N, S, Sr, H, F, A, B, P, Ax and D

5 Proposed System: NAERS

In this section, we discuss our proposed NAERS which stands for **N**ovelty **A**ware **E**motion **R**ecognition **S**ystem. Figure 1 shows a high-level diagram of different parts of this system.

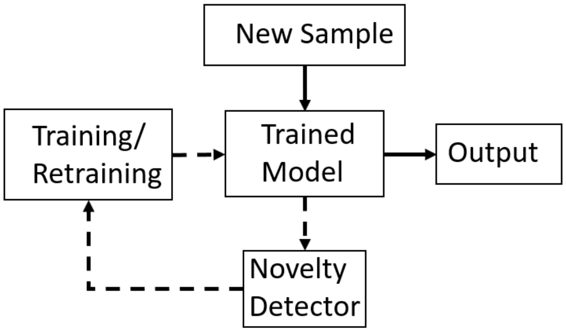

Figure 1: Proposed NAERS system architecture

5.1 Input

The input of this system is the samples of RGB frames with human subjects. This can be a static photo or a video frame. Live video can be also used for continuous recognition. In that case, we sample video frames in a suitable frequency such as 10 FPS (frames per second).

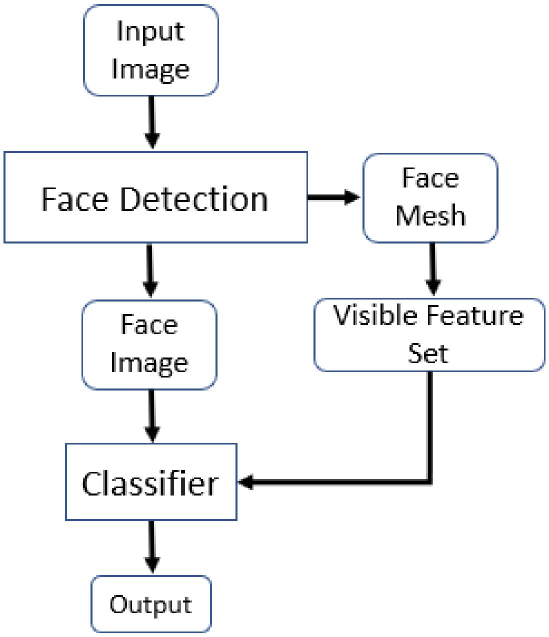

Figure 2: Proposed emotion recognition model architecture

5.2 Emotion Recognition (ER) Model

Face and Background Separation We use the facial image as our main source of the emotional information. There-

Table 2: Visible features

Feature Type	Feature Description
Width	Left eye Right eye Mouth
Height	Right eye Left eye Right eye Mouth
Distance	Left and right eyes Eyes to brows Eyes to mouth Eyes and nose Nose and mouth
Angle	Left eye with right eye and mouth Right Eye with left eye and mouth Mouth with both eyes Mouth with both eyes

fore, We separate the face from the background of the input frame. while the face is used for facial emotion recognition, the rest of the body and background are used for novelty detection (section 5.3). For facial area detection, we use the Blazeface face and body detection tool (Lugaresi et al. 2019). It can detect multiple landmarks on the face.

Visible Features Generation We create a face mesh using Blazeface face mesh generation tool (Lugaresi et al. 2019). From this mesh, we generate the features shown in the table 2. While shouting or laughing, people generally open their mouths resulting in a larger separation between the lips. Similarly, a surprised person will have big open eyes resulting in a bigger area. So these features convey meaningful information valuable to us. We also take the angle values between different face parts such as eyes and nose to capture facial distortion. They are normalized with face height to compensate for body size variations.

Deep Features Generation The emotion recognition system of NAERS uses the convolutional neural network to generate deep features from the input images. These deep features are then concatenated with the visible features (5.2) to generate the final feature vector for the classification model. For deep features generation, we experiment with three types of CNN:

- Regular CNN: We use two convolutional and two fully connected layers as shown in figure 3. Input images are scaled to 226 × 226. We use a filter size of 2 × 2 and a stride of 1. The downsampling layers used 2 × 2 max-pooling.
- Transfer Learning: We use the transfer learning approach in this type using Resnet-50 ((He et al. 2015)) trained with ImageNet. Resnet-50 is a 50 layers deep neural network.
- Ensemble CNN: In ensemble learning several primary models (weak learners) are trained individually on the data set. We use multiple regular CNN models as our weak learners. The model variations come from the variation of the initial weights. We used the average of the outputs from the model as the final output

We do not use the classification output provided by these deep feature generation models. Instead, we use the representation of the input samples at the output of the last layer before the linear layers as deep features.

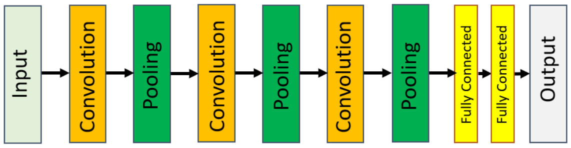

Figure 3: Proposed Convolutional Neural Network (CNN) architecture for deep features generation

Emotion Classifier Both the visible and deep feature sets are concatenated to form a final feature vector. We pass this feature vector through a 3-layer deep neural network (DNN) which provides class probabilities for the input sample. In the deep layers, we use the Relu activation function and the final layer output is passed through a softmax function to generate the probability values. The model is trained in the mini-batch gradient descent fashion. Different mini-batch sizes, selection of an optimizer from options such as Momentum, Adam and Nesterov, and learning rate are the hyperparameters of the model.

5.3 Novelty Detector

We propose a new algorithm for the novelty detector to detect novel and unexpected data samples. It relies on available multimodality of the data samples:

Additional modality As we discussed in section 2, human emotion can be identified from not only their face but also from posture and gesture extracted from the same input sample. For example, a confident person would appear relaxed while an afraid person would appear tensed and shrunk. For the input sample, we also analyze posture. If both facial and posture network output does not concur we tag that sample as a potential novelty. For posture based detection, we remove the facial area from the input. We use a kinematic human body model to represent the body posture and identify key body joints on the frame using the Blazepose tool. By following a similar idea we used in section 5.2, we create a visible feature set from different joint angles and distances. We use three CNN models (section 5.2) to generate deep features. The final feature set is created by concatenating visible and deep feature sets. The classifier is also a similar 3 layer deep neural network. For the brevity of the manuscript, we do not provide the details of the posture network here.

Context Context is important for novelty detection. A tiger in a forest is a normal sample but the same tiger in a busy city street indicates something is wrong. We create context from the background of the image. At first human body and face is removed from the input frame. Then we compare

Author	Model	Accuracy(%)
(Mollahosseini, Chan, and Mahoor 2016)	CNN	66.0
(Dhankhar 2019)	RESNET-50	67.2
(Renda et al. 2019)	Ensemble	71
(Gan, Chen, and Xu 2019)	Ensemble	73.73
This work	Ensemble CNN	75.8

Table 3: Performance comparison of different emotion recognition systems on FER-2013 dataset

Dataset	Accuracy(%)
CK+	97.5
FER-2013	75.8
AffectNets	66.5
CAER-S	72.4
FABO	96.1

Table 4: Performance of our emotion recognition model on different related datasets

sample background with the training background distribution using clustering techniques. If a significant difference in the background is detected we tag the input sample as a novelty.

5.4 Retraining

Retraining is done periodically and initiated when a sufficient number of novelty samples are encountered. During the retraining phase, we explore the clustering options to check whether a new class label is warranted or not. if a large number of novelty samples gather around together that indicates the requirement of a new class label. The most probable reason for that may be an important class label being missing from the training data. So we add this new class label. Right now re-labelling is done by human operators. We are exploring other options possibly without human involvement as future work. For the situation where novelty samples do not indicate new class label requirement, rather a mismatch between face and posture modalities, we update their class label with more precise manual labelling and increase their weight in the training sample distribution in Adaboost fashion.

6 Experimental Results

In this section, we discuss some of the experiments we have performed so far.

6.1 Performance of Emotion Recognition Model

We got 68%, 71.5% and 75.8% test accuracies for the three versions of our model with different CNN models on FER-2013 dataset. In addition, we compared our results with similar works on this dataset as shown in table 3. Our approach with the ensemble of CNN outperforms the other two models and other existing works. We attribute this improvement to the variance reduction achieved by the ensemble operation. For the rest of this manuscript, we only consider ensemble CNN while talking about our emotion recognition model.

Figure 4 shows the confusion matrix for our emotion recognition model for the FER-2013 dataset. Here actual emotion labels are plotted along the vertical while predicted labels are along the horizontal axis. An entry 'C_{ij}' of row 'i' and column 'j' represents the fraction of total samples having the true label of row 'i', and predicted label of column 'j'. The diagonal values are the accuracy of the class and a higher value is represented by a darker colour. We notice the happiness class has higher accuracy implying it is easier to detect. For fear, the accuracy is lower and that's because the system confuses fear with sadness. These two are closely related emotions that may appear together and people may actively try to conceal their fear with other similar emotion classes.

The results of the facial emotion recognition model on some other datasets are shown in table 4. Our model does well while comparing other works in FER-2013, AffectNet and CAER-S datasets. For the FABO dataset, we outperformed the accuracy of 76.4% reported by the authors of the dataset (**?**) by a wide margin. For the CK+ dataset, we found the best-reported result to be 98.57% accuracy (Kurup, Ajith, and Ramón 2019). Our accuracy of 98.5% is close to it.

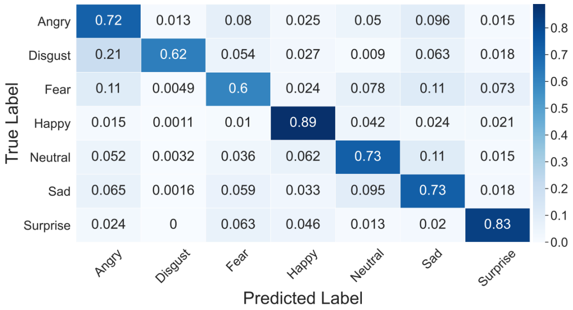

Figure 4: Confusion matrix of our emotion recognition model for FER-2013 dataset.

6.2 Novelty Detection Experiment

As mentioned earlier in section 5.3, for novelty detection we used posture as an additional modality and check the mismatch between facial based and posture based recognition. For training the posture emotion recognition model, We used several posture emotion recognition datasets such as FABO, Emotic and CAER-S.

In the FABO dataset, our posture model achieved 83.1% accuracy. Results from several other recent posture based work on FABO dataset are shown in table 5. We notice from the table that our model does a good job extracting the underlying emotion from the posture. From this table, we can see there is a mismatch between face and posture emotion recognition accuracies which means not all samples are getting similar output labels from both models. Further analysis of the results shows that there is around 12.8% sample that gets different class label output from face and posture recognition. Using our re-weighting and retraining algorithm we could reduce the mismatch close to 5.8%. Further reduction is hard to achieve and we attribute this fact to the idea that the face conveys more information than the posture.

Table 5: Posture emotion recognition accuracy comparison on FABO dataset

Author	Model	Face	Posture
(Barros et al. 2015)	CNN	72.07	57.8
(Chen et al. 2013)	SVM with the RBF kernel	66.5	66.7
(Gunes and Piccardi 2009)	Adaboost	35.2	73.2
This work	Ensemble CNN	92.7	83.1

6.3 Novelty Detection Cost

We measured the average time it takes to perform classification using face and posture modalities. We tested them on the datasets mentioned earlier on an Intel 2.8 GHz Core I7 laptop computer without any GPU access. We found facial emotion recognition requires less time than posture based recognition. From table 6 we can see that adding posture analysis to detect novelty significantly increases processing time. So this is a trade-off that can be decided based on the systems need to handle novelty.

Mode	Time
Face	1
Posture	1.5

Table 6: Average time taken by posture based model with respect to facial based model (values represents proportion).

7 Conclusion

In this paper, we presented a NAERS-a novelty aware emotion recognition system that deals with novelty in a continuous learning fashion. We showed some of the early experiments with promising results. Some of the components on the system are yet to be completed and we are focusing on those as the next step of this work.

8 Acknowledgement

This research is supported, in part, by the Defense Advanced Research Projects Agency (DARPA) and 3the Air Force Research Laboratory (AFRL) under the contract number W911NF2020003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the U.S. Government. We thank our team members on this project for all the discussions to develop this paper. Some of the ideas in this paper are based on our learning from the SAIL-ON meetings

References

Barros, P.; Jirak, D.; Weber, C.; and Wermter, S. 2015. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks*, 72: 140–151.

Chen, S.; Tian, Y.; Liu, Q.; and Metaxas, D. N. 2013. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, 31(2): 175–185.

Datta, S.; Sen, D.; and Balasubramanian, R. 2017. Integrating geometric and textural features for facial emotion classification using SVM frameworks. In *Proceedings of International Conference on Computer Vision and Image Processing*, 619–628. Springer.

Dhankhar, P. 2019. ResNet-50 and VGG-16 for recognizing facial emotions. *International Journal of Innovations in Engineering and Technology (IJIET)*, 13(4): 126–130.

Gan, Y.; Chen, J.; and Xu, L. 2019. Facial expression recognition boosted by soft label with a diverse ensemble. *Pattern Recognition Letters*, 125: 105–112.

Gunes, H.; and Piccardi, M. 2009. Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1): 64–84.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

Jadhav, R. S.; and Ghadekar, P. 2018. Content based facial emotion recognition model using machine learning algorithm. In *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, 1–5. IEEE.

Kurup, A. R.; Ajith, M.; and Ramón, M. M. 2019. Semi-supervised facial expression recognition using reduced spatial features and deep belief networks. *Neurocomputing*, 367: 188–197.

LeCun, Y.; Haffner, P.; Bottou, L.; and Bengio, Y. 1999. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, 319–345. Springer.

Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M. G.; Lee, J.; et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Mollahosseini, A.; Chan, D.; and Mahoor, M. H. 2016. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, 1–10. IEEE.

Patel, K.; Mehta, D.; Mistry, C.; Gupta, R.; Tanwar, S.; Kumar, N.; and Alazab, M. 2020. Facial sentiment analysis using AI techniques: state-of-the-art, taxonomies, and challenges. *IEEE Access*, 8: 90495–90519.

Renda, A.; Barsacchi, M.; Bechini, A.; and Marcelloni, F. 2019. Comparing ensemble strategies for deep learning: An application to facial expression recognition. *Expert Systems with Applications*, 136: 1–11.