2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia

Facial Expression Recognition on Video Data with Various Face Poses Using Deep Learning

Ayas Faikar Nafis
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ayas16@mhs.if.its.ac.id

Dini Adni Navastara
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
dini_navastara@if.its.ac.id

Anny Yuniarti
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
anny@if.its.ac.id

***Abstract*—Facial expressions in humans produce non-verbal communication to convey emotional states in humans; hence, they play an essential role in social interactions between humans. Along with the times, research on facial expression analysis has expanded to automatic facial expression recognition by computers. The facial expression recognition plays a vital role in human-computer interactions, monitoring human behavior, educational techniques, psychological, to sociable robots. In this study, the development of human facial expression recognition was carried out using a deep learning method called You Only Look Once (YOLO) based on Convolutional Neural Network (CNN). There are seven classes of facial expressions that can be recognized, namely angry, disgust, fear, happy, sadness, surprise, and neutral. The datasets used are video-based facial expression datasets such as CK+, IMED, and video data from 8 students of the Informatics Department, Institut Teknologi Sepuluh Nopember (ITS), with various face poses. Based on the experimental results, the best accuracy of the still image dataset is 94% on the CK+ dataset with channel three and learning rate 0.01. Moreover, the accuracy of video data with various face poses achieves 73%.**

Keywords— CNN, deep learning, Facial Expression Recognition, various face poses, video data, YOLO

I. INTRODUCTION

Humans are social creatures who can communicate with other humans in a verbal or non-verbal way [1]. Facial expression is a form of human communication in the form of non-verbal expression through the muscles on the face to convey emotional states in humans. It plays an important role in social interactions between humans [1].

The recognition of facial expressions has long been an interesting area of research. Research on facial expression analysis conducted by psychologists found six basic human expressions, namely angry, disgust, fear, happy, sad, and surprised. The analysis of facial expressions is a major research area of psychology, and many works and literature are published in this field [2].

However, along with the development of the era, research on facial expression analysis has penetrated automatic facial expression recognition by computers. Facial expression recognition plays an important role in human-computer interaction, monitoring of human behavior, educational techniques, psychological, and sociable robots [3]. Computers can learn like humans and detect human expression patterns with deep learning methods, which are part of machine learning [4].

Facial expression recognition with computers can be done with a deep learning-based object detection method to detect facial areas and classify the facial expressions. J. Li et al. developed facial expression recognition using the Faster R-CNN method using a dataset that has been taken from TV series or movies with various face poses [5]. J. Bao et al. also developed facial expression recognition using the Faster R-CNN based method using self-taken datasets with no explanation on how the environment or faces is positioned when the dataset has been taken [6]. However, Faster R-CNN is the object detection method with two stages, namely the detection stage and the classification stage, which are not efficient and can make the processing time longer [4].

One solution to overcome this problem is to use an object detection method that only performs one step for detection and classification. One of the most recently used deep learning methods for object detection and classification is You Only Look Once (YOLO). YOLO is an object detection method based on Convolutional Neural Network (CNN), which can produce fast and effective object detection and is one of YOLO's advantages over other object detection methods [4].

The YOLO method has been used in research to make sociable robots that can recognize the robot user's facial expressions so that the robot can know the user's emotions and respond appropriately to the robot user [3]. The study also conducted tests with data based on actual conditions with various lighting and facial positions. However, the YOLO method used is YOLO that has been trained only to detect faces called YOLO Faced, while facial expression recognition uses Ensemble CNN. So it can be said that it still uses two stages to recognize human facial expressions, which is inefficient.

In this study, the YOLO method, specifically YOLOv3, is used to recognize facial expressions based on video data with various face poses based on the real-life environment. The YOLO method is used because it is an object detection method that has one stage for detection and classification. There are seven classes of facial expressions that can be recognized, namely angry, disgust, fear, happy, sadness, surprise, and neutral. The dataset used is a video-based facial expression dataset such as The Extended Cohn-Kanade AU-Coded Facial Expression Database (CK+) [7], The Indonesian Mixed Emotion Dataset (IMED) [8], and video data from 8 students of the Informatics Department, Institut Teknologi Sepuluh Nopember (ITS). The video data has various face poses. YOLO can recognize facial expressions with different face poses based on real-life conditions such as facing the camera, turning right slightly, turning left slightly, breaking the head to the right, and breaking the head to the left. First, the dataset is preprocessed, and training is carried out using YOLO. Testing is done using testing data that will produce results in the form of a bounding box, a prediction of facial expression classes, and a confidence score. Evaluation is done by comparing the class of the predicted results with the ground truth class using measurements of accuracy, precision, recall, and F1-Score.

978-1-7281-1097-4/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: SRI SIVASUBRAMANIYA NADAR College of Engineering. Downloaded on November 05,2022 at 06:09:56 UTC from IEEE Xplore. Restrictions apply.

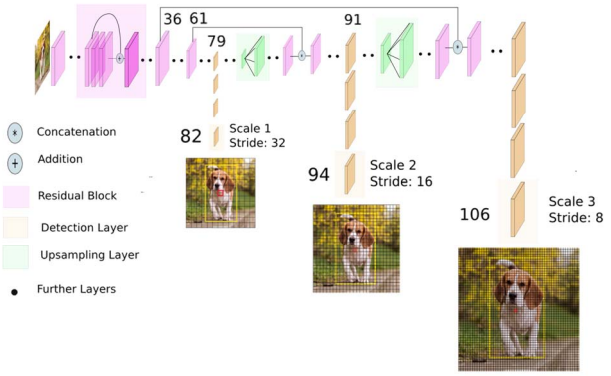


Fig. 1. YOLOv3 Architecture [10]

II. BASIC THEORY

A. *You Only Look Once (YOLO)*

YOLO (You Only Look Once) is the one-shot architecture that was successfully implemented to identify and classify objects simultaneously [4]. YOLOv2 or YOLO9000 is an updated version of the previously YOLO network. The YOLOv2 can run in a number of sizes, and the multi-scale training method provides a trade-off between speed and accuracy [9]. Furthermore, the last version of YOLOv3 came out in 2018. YOLOv3 has a new architecture for feature extraction called Darknet-53. YOLOv3 is the development of YOLOv2. Unlike YOLOv2, YOLOv3 architecture is based on Darknet-53, which uses 53 convolutional layers for feature extraction.

YOLOv3 is used in this study. Compared with YOLOv2, YOLOv3 has many advancements. First, YOLOv3 predicts bounding boxes at three different scales, and the scale is given by taking a sample of the input image dimensions of 32, 16, and 8, respectively. This scale is used to detect large, medium, and small objects. As a result, YOLOv3 predicts ten times the number of bounding boxes predicted by YOLOv2. Second, the loss function is changed from squared error to cross-entropy error. In other words, the confidence object for each bounding box is predicted using logistic regression. Finally, to detect objects, YOLOv3 performs a multi-label classification . Hence, the softmax function was replaced by an independent logistic classifier.

YOLOv3 has 75 Convolutional Layers, 23 Shortcut Layers, 3 YOLO Layers, 4 Route Layers, and 2 Upsample Layers, making YOLOv3 have 107 layers as in Fig. 1 [10]. Shortcut layers are layers that add feature map layers to the previous nth layer. YOLO layer is a layer that issues object predictions with a bounding box. The route layer is the layer that takes the feature map from the previous nth layer. The Upsample layer is a layer that can enlarge the feature map on the previous layer.

The way YOLOv3 works is almost the same as the previous YOLO, namely dividing the image into S × S grids, then each grid will detect objects around it. However, YOLOv3 uses an Anchor Box or Prior Box, which is a bounding box that has been defined at the beginning of various sizes. YOLOv3 uses three Anchor Boxes for each scale to detect objects in each cell so that YOLOv3 has a total of nine Anchor Boxes [11].

B. *The Extended Cohn-Kanade AU-Coded Facial Expression Database (CK+)*

The Extended Cohn-Kanade AU-Coded Facial Expression Database (CK+) is a collection of human face images with a total of 593 sequential images of 123 subjects with eight classes of expression including neutral with 123 subjects, angry with 45 subjects, happy with 69 subjects, sadness with 28 subjects, disgust with 59 subjects, surprise by 83 subjects, fear of 25 subjects, and contempt with 18 subjects. The image in this dataset has an image size of 640 × 490 or 640 × 480 pixels. The example image for each class of facial expressions in the CK+ dataset can be seen in Fig. 2 [7].

Fig. 2. The Extended Cohn-Kanade AU-Coded Facial Expression Database (CK+) [7]

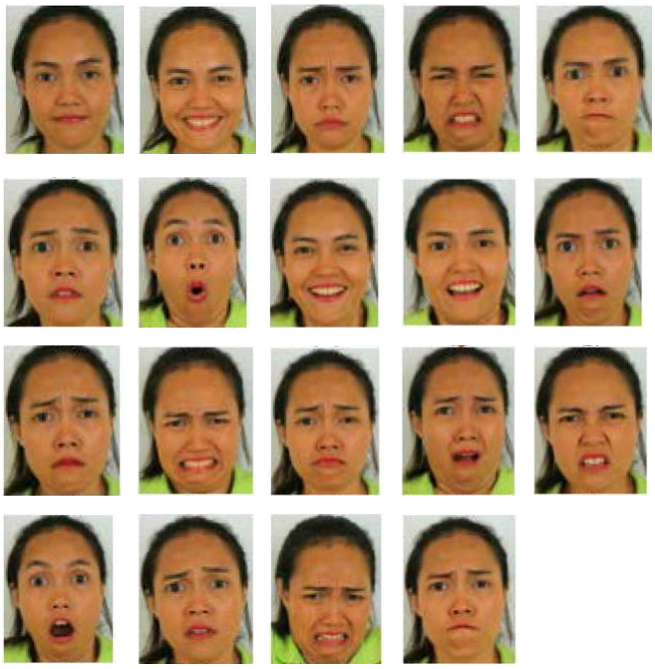


Fig. 3. The Indonesian Mixed Emotion Dataset (IMED) [8]

C. *The Indonesian Mixed Emotion Dataset (IMED)*

The Indonesian Mixed Emotion Dataset (IMED) is a video-based dataset of facial expressions from Universitas Indonesia. IMED has 19 categories of emotions carried out by 15 subjects, all of whom are Indonesians with various ethnicities, including Javanese, Sundanese, Malay, Batak, Minang, and Manado. Subjects were 60% female and 40% male ranging in age from 17 to 32 years old who demonstrated basic and mixed emotion classes in the video.

The IMED dataset has emotional or class categories, including neutral, happy, sadness, disgust, angry, fear, surprise, pleasant-surprised, happy-disgust, sadness-surprise, sadness-surprise, sadness-disgust, fear-surprise, disgust-surprise, angry-surprise, fear-angry, fear-disgust, and disgust. An example image for each class of facial expressions on the IMED dataset can be seen in Fig. 3 [8].

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on November 05,2022 at 06:09:56 UTC from IEEE Xplore. Restrictions apply.

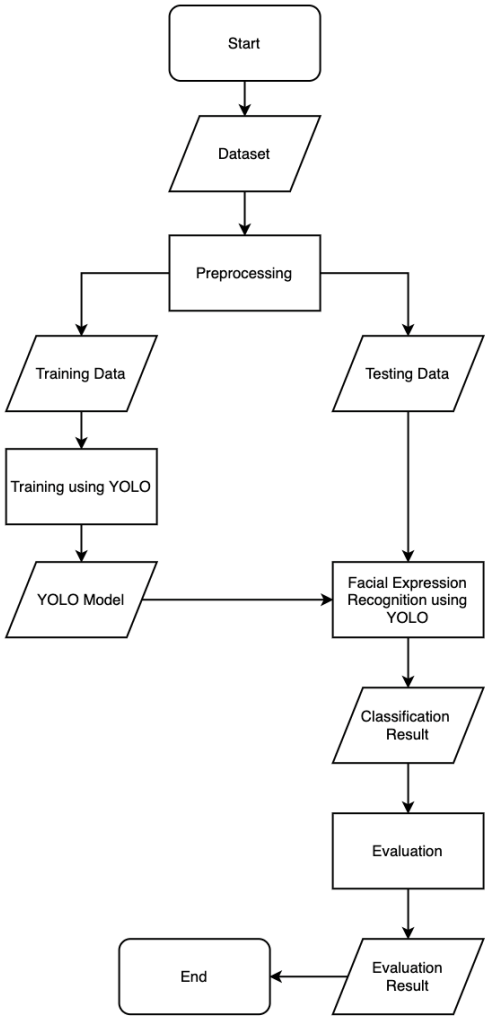


Fig. 6. System Flow Diagram

III. METHODS

The facial expression recognition system built has the main process, namely preprocessing, training using YOLO, facial expression recognition using YOLO, and evaluation. The system flow diagram is shown in Fig. 4.

A. Dataset

The dataset used to conduct training at YOLO is The Extended Cohn-Kanade Dataset (CK+), The Indonesian Mixed Emotion Dataset (IMED), and facial expression video data from eight students of Informatics Department, ITS. Each subject has seven videos which each video has one facial expression. This facial expression video has various face poses such as facing the camera, turning right slightly, turning left slightly, breaking the head to the right, and breaking the head to the left. The examples of happy expression on video data with various face poses can be seen in Fig. 5. There are seven categories of emotions or classes used in this study, namely happy, sadness, angry, disgust, fear, surprise, and neutral.

B. Preprocessing

The preprocessing stage is useful for separating the dataset into training data and testing data with the same number of images for each expression class so that the model does not have more bias towards one of the expression classes. This stage also equalizes the image size of each dataset to a certain size so that the training process can run

Fig. 4. The examples of happy expression on video with various face poses

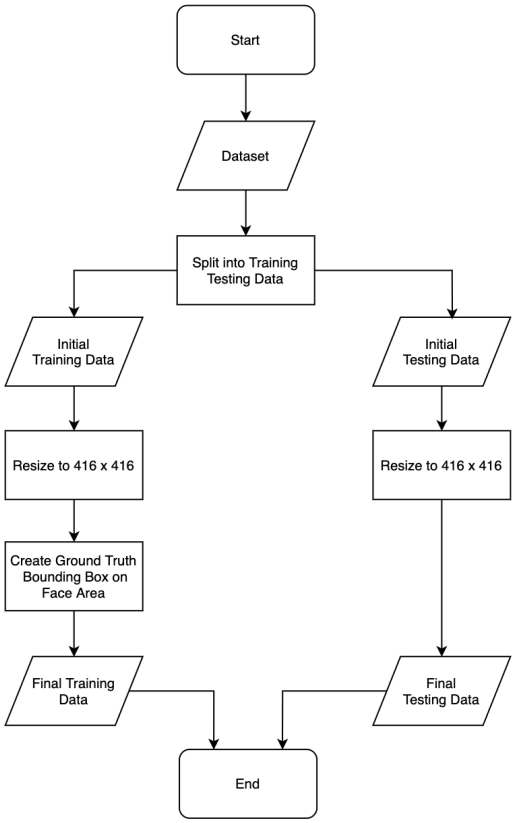


Fig. 5. Preprocessing Flow Diagram

efficiently. The preprocessing stage was carried out by dividing the CK+, IMED, and video data into training and testing data with different subjects. The number of subjects for training and testing for each expression class is determined so that the amount of data between each expression class is even. Some subjects in the CK+ dataset do not have all the expression classes and may have one or more expression classes. For the CK+ dataset, 25 random subjects are in the training data, and ten other subjects are in the testing data because the *fear* class has 25 subjects that is the least number of subjects in the CK+ dataset besides the contempt class unused in this study. As for IMED, ten random subjects are in the training data, and five subjects are in the testing data. For video data, five random subjects are in the training data, and three other subjects are in the testing data. All subjects in each expression class may differ from each other.

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on November 05,2022 at 06:09:56 UTC from IEEE Xplore. Restrictions apply.

The CK+ and IMED datasets are video-based, so each expression class in each subject consists of a sequence of facial images with expressions. The CK+ dataset has a facial image sequence starting from neutral expressions to peak expressions. For the neutral expression class, the first three images from the expression class for each subject are taken.

For other expressions except for fear and sadness, the last three images from the expression class are taken for each subject. For the *fear* and *sadness* classes, which have 25 subjects and 28 subjects respectively, the training data use the last three images from expression class, and the testing data use three images before the last three images from expression class. Whereas the IMED dataset has a facial image sequence starting from a neutral expression, a peak expression, and returning to a neutral expression so that five images are taken at the top of the expression for each class. In the video data, three images are taken for each face pose for training data.

After that, the image in the training and testing data is resized to 416 × 416 pixels. Each image in the training data is created with a bounding box ground truth in the face area using Multi-task Cascaded Convolutional Networks (MTCNN) [12] to help the creation of a ground truth bounding box. After that, if a ground truth-bounding box is not suitable or the face location is not detected, it is necessary to manually repair or create a ground truth bounding box. Ground truth bounding boxes are used to conduct training YOLO. The flow chart for the preprocessing stage can be seen in Fig. 6. Preprocessing CK+, IMED, and video data datasets that have been preprocessed are combined into one for training.

C. Training using YOLO

The training stage was carried out using a pre-trained model from YOLOv3, which was carried out on the ImageNet dataset and using training data that had gone through the preprocessing stage. The training phase is carried out using built-in hyperparameters with batch size 64, subdivision 16, and max batches 6000. Max batches function like iterations in the training stage. YOLO will conduct training in batches containing 64 images and divided into 16 subdivisions or mini-batches for training. After one batch has finished training, it is followed by a batch training after that in the next iteration. The hyperparameter channel and learning rate can be changed according to the test scenario at the training stage. At the end of each iteration, there is a model testing process to determine how well the model is being trained. YOLO's initial hyperparameter specifications can be seen in TABLE I.

D. Facial Expression Recognition using YOLO

After the training phase is complete, YOLO will produce a model that has been trained and ready for facial expression recognition testing. YOLO was tested using data testing that had been carried out in the preprocessing stage. A hyperparameter that can be changed at this stage is the channel, which has the same initial values during training. YOLO will detect and classify facial expression classes on each test image using the YOLO model that has been trained and generate a bounding box containing the position, classification results, and confidence score.

E. Evaluation

The classification results that have been obtained through the detection and classification stages will be used to evaluate the YOLO model that has been trained. The evaluation stage serves to measure how well the model can correctly predict facial expression classes on data that the model has never seen. At the end of the test, calculations will be carried out to get accuracy, precision, recall, and F1-Score values.

IV. Experimental Result

The testing process is useful for finding parameters that produce the most optimal model performance. The right parameters will give better results during the testing process. The best results from a test scenario will be used for the next test scenario. In this section, there are test scenarios and test results, as follow:

A. YOLO Parameter Exploration

The YOLO parameter exploration aims to find the best parameter values that can produce the best performance model. The scenario was tested by changing the channel and learning rate parameter from the initial hyperparameter in TABLE I. This scenario was tested on testing data with 385 facial expression images. The scenario tested at YOLO is to change the channel parameter and learning rate from the original hyperparameter.

First, YOLO was tested by varying the input color channel parameters. Color channel parameter exploration is useful for knowing how well the model is when training with different color channels, in this case with a value of 1 (grayscale) and 3 (RGB). Another hyperparameter, such as learning rate, is using the value from the initial hyperparameter. The results of channel parameter replacement tests on the YOLO architecture can be seen in TABLE II.

The results of testing the replacement of parameters at YOLO by varying the channel or color channel's parameters show that YOLO with three channels gets an accuracy of 87%, which is better than one channel, which has an accuracy of 86%. The 3-channel setting performs better than

TABLE I. Initial YOLO Hyperparameter Specification

Information	Specification
Size	416×416
Max_batches (iteration)	6000
Batch	64
Subdivisions	16
Color channel	3 (RGB)
Learning Rate	0.01

TABLE II. Channel Variation Test Result

Channel	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	86	87	86	85
3	**87**	**88**	**87**	**87**

TABLE III. Learning Rate Variation Test Result

Learning Rate	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
0.1	86	89	86	86
0.01	**87**	**88**	**87**	**87**
0.001	86	88	86	86

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on November 05,2022 at 06:09:56 UTC from IEEE Xplore. Restrictions apply.

the 1-channel because the 3-channel training model can learn the three channels' features so that when testing, the model can see facial expressions more clearly.

Furthermore, YOLO was tested by varying the learning rate values that can produce the best models. The initial value of the learning rate in this study is 0.01, with other learning rate values for comparison being 0.1 and 0.001. The channel parameter used in this test scenario is the channel that gets the best results from the previous test, namely channel 3. The results of testing the learning rate parameter replacement on the YOLO architecture can be seen in TABLE III. The parameter replacement test results by varying the learning rate parameters show that YOLO with a learning rate of 0.01 gets an accuracy of 87%, which is better than the 0.1 and 0.001 learning rate. Therefore, with a learning rate of 0.01, the model can make optimal weight corrections during the training process.

B. Scenario-Based on The Dataset

The test scenario based on the dataset aims to find out which dataset has the best performance when tested with a model that has been trained. In this scenario, the YOLO model is tested with testing data from two different datasets, namely The Extended Cohn-Kanade AU-Coded Facial Expression Database (CK+), which has 210 facial expression images with 30 images in each expression class, and The Indonesian Mixed Emotion Dataset (IMED), which has 175 images of facial expressions with 25 images in each expression class. The model used for testing has been trained using the best channel and learning rate parameter from the YOLO parameter exploration. The test results based on the dataset can be seen in TABLE IV. The test results based on the dataset show that the CK+ dataset has higher accuracy than the IMED with an accuracy of 94%.

C. Scenario-Based on The Video Data

The test scenario on video data aims to test the YOLO model's performance with data that has never been seen before and taken in uncontrolled environmental conditions. Testing on video data was carried out with video data from 8 students of the Informatics Department, ITS, with various face poses.

The evaluation is carried out by calculating the majority vote of each video. The video data was tested using two YOLO models, which were carried out by training with different datasets. The first model or Model 1 is trained with the CK+ and IMED dataset, while the second model or Model 2 is trained with the CK+, IMED, and video data. The results can be seen in TABLE V. The results of testing facial expression recognition from video data with YOLO and using a majority vote shows that 73% of videos were classified correctly or 41 videos from a total of 56 videos using the Model 2. Almost two times better than the Model 1 dataset that can correctly predict 37% or 21 videos from a total of 56 videos.

Video data is taken in an uncontrolled environment and varies with each video, unlike the dataset taken in controlled environments such as lighting, the distance between faces and cameras, and background conditions. Some subjects are less able to express themselves, as in Fig. 7(c). For example, the surprise class generally has the characteristic of being wide-eyed and the mouth wide open, as in Fig. 7(a) and Fig. 7(b). The video data also has different face poses and facial expressions with one another have similarities as in Fig. 8.

TABLE IV. SCENARIO BASED ON DATASET RESULT

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CK+	**94**	**95**	**94**	**94**
IMED	79	80	79	77

TABLE V. SCENARIO BASED ON THE VIDEO DATA RESULT

Class	Total Subject Predicted Correctly		Subject Total
	Model 1	Model 2	
Neutral	3	7	8
Angry	2	6	8
Disgust	6	5	8
Fear	0	5	8
Happy	6	7	8
Sadness	0	6	8
Surprise	4	5	8
Total	21	41	56

TABLE VI. METHOD COMPARISON RESULT

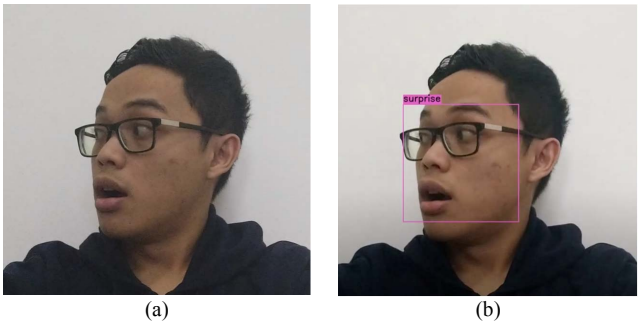
Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (seconds)
YOLOv3	**87**	**88**	**87**	**87**	**44.85**
Faster R-CNN	72	76	72	71	456.34

(a) (b) (c)

Fig. 7. The examples of comparison of surprise class in dataset (a) CK+, (b) IMED, and (c) Video Data

(a) (b)

Fig. 8. Examples of Similar Facial Expressions in Video Data. (a) Actual: Disgust, Predicted: Happy, (b) Actual and Predicted: Happy



(a) (b)

Fig. 9. Example of Subject 3 Surprised Class Turns Right. (a) Face not detected on Model 1, (b) Face detected on Model 2

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on November 05,2022 at 06:09:56 UTC from IEEE Xplore. Restrictions apply.

Fig. 8(a) is the disgust class, which is predicted to be the happy class, while Fig. 8(b) is the happy class that is predicted as a happy class.

The comparison of the results of the model trained with Model 1 and Model 2 can be seen in Fig. 9. Fig. 9 shows that the trained model without using video data cannot detect faces, as in Fig. 9(a) while the model trained with additional video data can detect faces and classify facial expressions as in Fig. 9(b).

D. Method Comparison

The method comparison test scenario aims to compare the YOLO method's performance with other object detection methods. In this test scenario, YOLO is compared with another object detection method called Faster R-CNN [13] with Google's Inception v2 preset. The configuration parameters used in this Faster R-CNN has been adjusted to the best parameters used for YOLO. This scenario was tested on testing data with 385 images of facial expressions. The results of the comparison test between the YOLO and Faster R-CNN can be seen in TABLE VI. The comparison between YOLO and Faster R-CNN shows that YOLO has a higher accuracy of 87% and a faster running time of 44.85 seconds on the testing data with 385 facial expression images. This is because Faster R-CNN creates a feature map with CNN first, then looks for the Region of Interest (ROI), and each ROI is classified by Region Proposal Network (RPN), while YOLO only performs CNN once to detect and classify objects.

V. Conclusion

The YOLO method, which uses only one stage for detection and classification, is enough to produce good results. Based on the test results, the YOLO model produced the best accuracy of 94% on the CK+ dataset with an average accuracy of 87% with channel 3 and learning rate 0.01. YOLO is also proven to be better and faster than Faster R-CNN with an accuracy of 87% and running time of 44.85 seconds. Based on video data with various face poses, the model can correctly predict 73% of facial expression videos or 41 of 56 videos. The YOLO method is pretty good for classifying real-life data videos with various face poses with different lighting and background. Suggestions for further works are the subject of the dataset must have explicit facial expressions and can be distinguished from one facial expression to another.

Acknowledgment

This study was funded by Directorate of Research and Community Service (Direktorat Riset dan Pengabdian Masyarakat, DRPM) Institut Teknologi Sepuluh Nopember (ITS) Surabaya with the grant number of 1665/PKS/ITS/2020.

References

[1] I. M. Revina and W. R. S. Emmanuel, "A Survey on Human Face Expression Recognition Techniques," *J. King Saud Univ. - Comput. Inf. Sci.*, 2018, doi: 10.1016/j.jksuci.2018.09.002.

[2] N. B. Kar, K. S. Babu, A. K. Sangaiah, and S. Bakshi, "Face expression recognition system based on ripplet transform type II and least square SVM," *Multimed. Tools Appl.*, vol. 78, no. 4, pp. 4789–4812, 2019, doi: 10.1007/s11042-017-5485-0.

[3] N. K. Benamara *et al.*, "Real-Time Emotional Recognition for Sociable Robotics Based on Deep Neural Networks Ensemble," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11486 LNCS, 2019, pp. 171–180.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 27, no. 3, pp. 779–788, doi: 10.1109/CVPR.2016.91.

[5] J. Li *et al.*, "Facial Expression Recognition with Faster R-CNN," *Procedia Comput. Sci.*, vol. 107, no. Icict, pp. 135–140, 2017, doi: 10.1016/j.procs.2017.03.069.

[6] J. Bao, S. Wei, J. Lv, and W. Zhang, "Optimized faster-RCNN in real-time facial expression classification," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 790, no. 1, p. 012148, doi: 10.1088/1757-899X/790/1/012148.

[7] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work. CVPRW 2010*, no. July, pp. 94–101, 2010, doi: 10.1109/CVPRW.2010.5543262.

[8] D. Y. Liliana, T. Basaruddin, and I. I. D. Oriza, "The Indonesian Mixed Emotion Dataset (IMED): A facial expression dataset for mixed emotion recognition," in *ACM International Conference Proceeding Series*, 2018, pp. 56–60, doi: 10.1145/3293663.3293671.

[9] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6517–6525, 2017, doi: 10.1109/CVPR.2017.690.

[10] A. Kathuria, "What's new in YOLO v3? – Towards Data Science," *Medium*, 2018. [Online]. Available: https://towardsdatascience.com/yolo-v3-object-detection-53fb7d3bfe6b. [Accessed: 22-Jun-2020].

[11] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Apr. 2018.

[12] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, 2016, doi: 10.1109/LSP.2016.2603342.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks."

Authorized licensed use limited to: Sri Sivasubramanya Nadar College of Engineering. Downloaded on November 05,2022 at 06:09:56 UTC from IEEE Xplore. Restrictions apply.