Bulletin of Electrical Engineering and Informatics
Vol. 11, No. 1, February 2022, pp. 213~220
ISSN: 2302-9285, DOI: 10.11591/eei.v11i1.3542

Explicit kissing scene detection in cartoon using convolutional long short-term memory

Muhammad Arif Haikal Muhammad Fadzli[1], Mohd Fadzil Abu Hassan[1,2], Norazlin Ibrahim[1,2]
[1]Universiti Kuala Lumpur, Malaysia France Institute, Bangi, Selangor, Malaysia
[2]UniKL Robotics and Industrial Automation Center (URIAC), UniKL MFI, Bangi, Selangor, Malaysia

Article Info

Article history:

Received Aug 17, 2021
Revised Nov 19, 2021
Accepted Jan 6, 2022

Keywords:

Darknet-19
Explicit cartoon
Kissing scene filter
Long short-term memory

ABSTRACT

The main concern of this study is due to certain cartoon content consisting of explicit scenes such as kissing, sex, violence. That are somehow not suitable for kids and may contradict to some religions and cultures. There are some reasons the film industry does not expel the kissing scene in a cartoon movie. It is categorized as a romance sequence and love scene. These could be a double-edged weapon that will ruin an individual’s childhood through excessive exposure to explicit content. This paper proposes a deep learning-based classifier to detect the kissing scene in the cartoon by using Darknet-19 for frame-level feature extraction, while the feature aggregation in the temporal domain is using convolutional long short-term memory (conv-LSTM). This paper also has discussed a few steps related to evaluation and analysis regarding the performance of the models. Extensive experiments prove that the proposed system provides excellent results of 96.43% accuracy to detect the kissing scene in the cartoon. Due to high accuracy performance, the model is suitable to be a kissing scene filter feature in a digital video player that may able to decrease the excessive exposure to explicit content for kids.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Mohd Fadzil Abu Hassan
Industrial Automation Section, Universiti Kuala Lumpur, Malaysia France Institute
Jln. Teras Jernang, 43650 Bandar Baru Bangi, Selangor, Malaysia
Email: fadzil@unikl.edu.my

1. INTRODUCTION

The internet has transformed unnecessary materials consumption for children. The internet has changed the fundamental relationship between the individual and the ‘things’ on the internet, allowing access to an endless supply of free and diverse material. Online materials such as videos, songs, and books are available 24/7 and accessible from virtually anywhere with an internet connection. With remote learning moving into the long term, experts say the mental and emotional impacts of that shift are likely to be challenging especially for preschool students and parents.

Nowadays, one of the common sources for distance learning for children is cartoon video. However, the cartoon is one of the influential factors that have a significant impact on a person's childhood [1]. The average of children with the facility of television at his home watches approximately 18,000 hours from kindergarten to high school [2]. Furthermore, the environment in which children grew up has a massive impact on their way of thinking. More worrisome is certain cartoons’ content consists of explicit scenes including kissing parts that are somehow difficult to expel for some reason. These could be a double-edged weapon that will ruin an individual’s childhood through excessive exposure to explicit content.

As reported by the Malaysia communications and multimedia commission [3], the term "indecent content" refers to material that is offensive, morally inappropriate, or violates the accepted behavior standards such as nudity and sex. They advise a special consideration must be given to the content created for children. Thus, they are providing a practical and commercially feasible guideline for the provision of content through self-regulation by the filmmakers. Currently, the filtering method for an explicit scene is to manually screen throughout the movie. Moreover, the film industry does not expel the kissing part include cartoon movies because kissing is just a romance sequence that has been symbolized for love and categorized as 'U' (unrestricted) category [4], which the content may be suited for children view.

Influence of cartoon towards children's behaviour. Every child's life is incomplete without the presence of cartoons. Several generations of children have grown up watching animated films since the invention of cartoon films over the past century. Statistics have shown that 2-5 years old children watched cartoons for 32 hours a week, while 6-7 years old children watched cartoons 28 hours a week [2]. According to the study, 71% of 8-18-year-old children had a television in their room, 53% of 7-12-year-old children had no parental monitoring, and 51% of homes television was turned on most of the time.

There are three factors in child brain development. First, thinking and imagination affect the functionality of the brain until the age of 12. Second, the surrounding experience influences the brain function, and lastly, early mind setting where the children pattern of future action could be predicted. Habib and Soliman [2] stated that the cartoon itself contained violence and sexual content. As a result, it may have an impact on children's brain development or, in the worst-case scenario, negatively impact children's behavior.

Machine learning and performance evaluation**.** There are a lot of methods to detect action and behavior using machine vision and artificial intelligence. Sudhakaran and Lanz [5] said that the best method to detect and identify the behavior of humans is by using the convolutional neural network (CNN). It extracts the frame-level features from a video and then aggregates them using a variant of the short-term long memory model. The convolutional long short-term memory (conv-LSTM) can capture the localized spatial-temporal features that enable the analysis of local motion taking place in the violent video. Núñez *et al.* [6] proposed a deep learning-based approach for temporal 3D pose recognition problems based on the combination of both algorithms. Instead of image data, multiple analog features extracted from tri-axial accelerometer and gyroscope sensors were used to differentiate five activities of daily living by using LSTM [7].

Another study in [8] suggested a scene detector specifically on the kissing scene in a movie. Ziai [8] proposed a prediction system that consists of two phases. The first phase consists of a binary classifier that predicts a binary label, i.e. kissing or non-kissing scenes and the second phase is to aggregate the binary labels for contiguous non-overlapping segments into a set of kissing scenes. The project was experimented with a variety of 2D and 3D convolutional architectures such as ResNet, DenseNet, and VGG to develop a highly accurate kissing detector. The performance of the conv-LSTM model generated 97% accuracy, while the standalone CNN model can perform 95% of accuracy. The other high-performance pre-trained CNN models such as densnet, alexnet, resnet, VGG, squeezenet, googlenet were proposed by previous researchers to solve various problem domains [9]-[12].

From a previous study in [13], the multi-task CNN model for attribute prediction was used and the study used two categories of the dataset: the clothing attributes dataset and the AwA dataset. The clothing attributes dataset consists of 1856 images with 23 binary attributes that have been annotated; while the AwA dataset is the animal with attributes dataset consisting of 30475 images of 50 animal classes. The annotation was made by the animal's class level and provided 85 binary attributes for each class. The dataset has nine groups. Another research on violence detection shows that the dataset's specification used consists of 50 frames (720x536 pixels) and it was manually labeled as fight or non-fight scenes [14]. Then, the research about online deep learning for action recognition also seems to use a lot of datasets consisting of 720x480 pixels of resolution with 10 frames per second. Somehow, the dataset also undergoing downgrading due to fit the frame patches [15]. However, the higher resolution and a greater number of samples will give better performance of the trained model. Nevertheless, a high-performance computing system should be considered. There are a lot of evaluation methods that have been used in machine learning such as confusion matrix and confidence interval receiver operating characteristic (ROC). A study of anomaly detection in surveillance video used the ROC for evaluating model performance [16], [17]. However, the confusion matrix method is the most commonly used in analyzing the classifier performance [12], [18], [19].

This study aims to build a machine learning-based classifier for kissing scene detection in a video image sequence. The specific objectives of this project are to develop a machine learning-based classifier that can detect human-cartoon kissing scenes in the video frame and evaluate the classification performance in terms of detection accuracy. The proposed model employs the convolutional neural network for frame-level feature extraction, while the long short-term memory (LSTM) is used for feature aggregation in the temporal domain. The rest of this paper is organized as: section 2 describes the proposed methodology in detail. While the results and discussion are presented in section 3. Finally, section 4 presents the conclusion and future work of this paper.

2. RESEARCH METHOD

There are three main processes in the development of the explicit kissing scene detector: i) data collection and preparation for the kissing dataset, ii) training the kissing scene detection model by using CNN, and iii) evaluating detection performances. A detail of these processes is presented in the following subsections:

2.1. Kissing dataset preparation

Datasets are a crucial part of the machine learning process. It determines the quality of machine learning. In this study, the video clip dataset had been collected from various sources such as YouTube and movies. It had gone through some clipping and editing processes. The supervised machine learning-based conv-LSTM had been trained with 300 cartoon clips with 30 frames/second. The dataset consists of kissing and non-kissing scenes with a duration range of three seconds. Two frame sizes are selected: 80x60 pixels and 320x240 pixels. Figure 1 shows the clipping process of a kissing scene in a video.

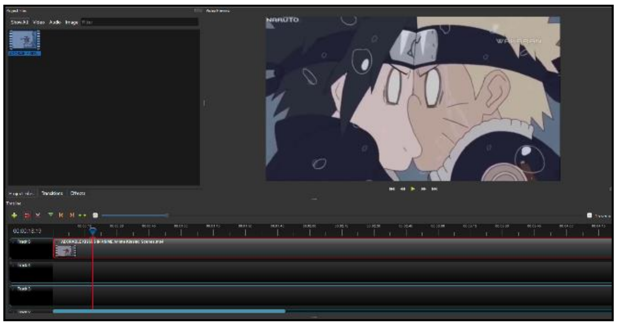


Figure 1. Clipping process using freeware open shot editor [20]

From 22 video samples of kissing scenes, 39 cartoon clips had been extracted which required about 2429.91 MB total of memory size and consist of 1561 seconds total of time duration. The clips must have a kissing scene within three seconds. Whereas 13 video samples of non-kissing scenes are used with 241 clips had been extracted from diverse types of actions such as cooking, laughing, crying, and swimming. All the clips required about 1829.9 MB total of memory size and consist of 1170 seconds total of the time duration. The number of clips is set to be balanced amongst the action categories to prevent bias. The dataset consists of two video clips: kissing and non-kissing. Then, it will be randomly split into training and testing sets. Part of the training set will be divided into validation sets; to understand the behavior of the model and generalizability on unseen data and to produce a completely unbiased estimation of model performance. Figure 2 shows the dataset distribution for training and evaluation.

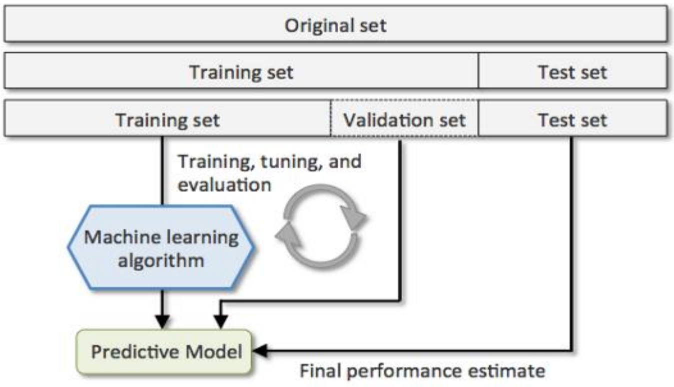

Original set
Training set
Test set
Training set
Validation set
Test set
Training, tuning, and evaluation
Machine learning algorithm
Predictive Model
Final performance estimate

Figure 2. Dataset distribution

2.2. Kissing detection model: conv-LSTM

The suggested study's goal was to create a trainable deep neural network model for classifying the kissing and non-kissing scenes in a video frame. The system should be able to identify the locations of the subject/individuals in the video cartoon scene and able to extract the motion (kissing action) features changes over time. The CNN promises to generate a good representation of each video frame, while a recurrent neural network (RNN) is required to encode the temporal changes. Since we are focusing on the encoding of spatial and temporal dimensions; thus, the conv-LSTM will be an appropriate option in this study, which may result in a better analysis of video representation. Therefore, for this dedicated video scene detection application, the pre-trained Darknet-19 model and the LSTM models were used to represent the CNN and RNN network; respectively.

Darknet-19 is made up of 19 convolutional layers, five max-pooling layers, and a soft-max layer for classifying objects. This model is pre-trained on the ImageNet database and is considered as fast execution with good classification performance in object detection which is very important for predicting in real-time [21]. Several studies demonstrated that the model has higher generalization and performs well in object and action recognition [22], [23]. Figure 3 illustrates the conv-LSTM architecture.

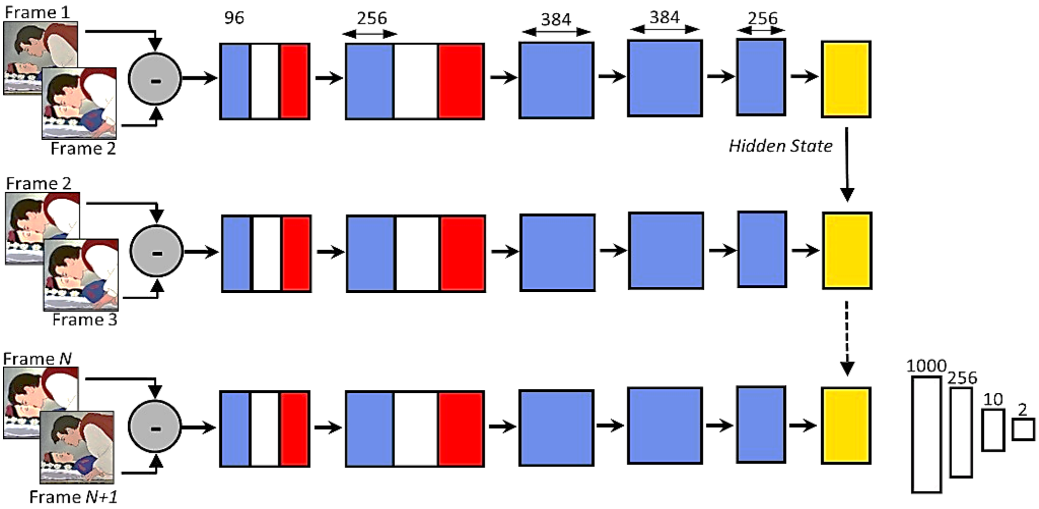


Figure 3. The model is made up of layers that alternate between convolutional (blue), normalization (white), and pooling (red), the final hidden state of conv-LSTM (yellow) is used for kissing and non-kissing scenes classification

The model's network consists of a CNN which comprises a set of convolutional layers followed by max-pooling operations in series for collecting discriminant features. Then, the CNN model is cascaded to LSTM for encoding the frame-level changes that distinguish kissing and non-kissing scenes in the video. The convolutional layers are trained to extract hierarchical information from video frames before it is aggregated to the LSTM layer. The convolutional gates of the conv-LSTM have been trained to represent temporal changes in a local region. As a result, the whole network can encode the localized spatiotemporal characteristics.

Instead of using the raw video frames only as input, the network uses the distribution of apparent velocities of movement of brightness pattern in an image between two consecutive frames as input. As a result, the network is forced to represent the changes occurring in a set of consecutive frames. This approach had been proposed by Dong *et al*. [24] of using optical flow features as input to a neural network for action detection.

2.3. Evaluation and analysis method

There is two evaluation method that has been applied throughout the project which is model fitting and confusion matrix. Model fitting is used to analyze the performance of the model is it robust, while the confusion matrix is used to see the accuracy and precision of the model to detect the labels. There are three-level to determine model fitness: underfitting, fitting, and overfitting. However, to get a better result, the experiment will just be conducted by taking the model that has a good fit only. To calculate the performance of the model, the confusion matrix method has been applied. Figure 4 shows the confusion matrix that has been applied in terms of two categories or classes which are kissing (positive label) and non-kissing (negative label). True positive (TP) explains that the actual value is a true state, and the prediction is positive [25].

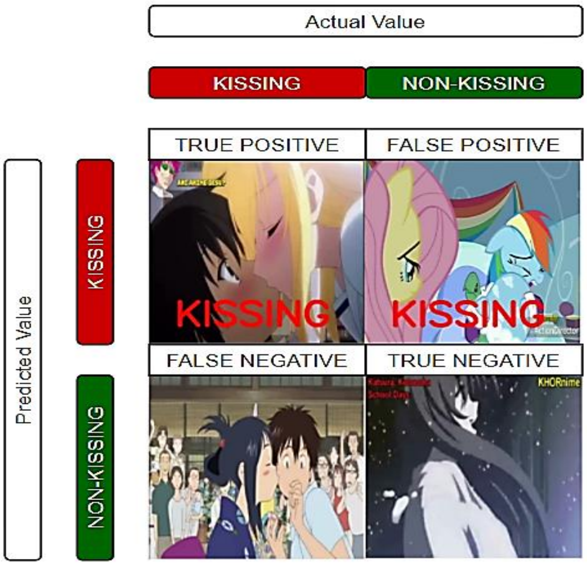


Figure 4. Confusion matrix for kissing and non-kissing detection

In the case of the kissing, the frame is predicted as a kissing scene, thus, it is TP. True negative (TN) explains that the actual value is a true state, and the prediction is negative (non-kissing label). From the case, the non-kissing clip has been deployed and the system predicted that is non-kissing, so it means TN. False positive (FP) or as type-1 error explains the actual value is a false state and the prediction is positive. From the case, the non-kissing clip has been deployed and the system predicted as positive, thus it is FP. False negative (FN) or as type-2 error explains the actual value is a false state and the prediction is negative. From the case, the kissing clip has been deployed and the system predicted as negative, so it means FN. For the accuracy calculation towards binary classifiers, the performance was made by calculating the ratio of true prediction (TP and TN) with all the predictions that have been made.

3. RESULTS AND DISCUSSION

This chapter will discuss the result of the experiment that has been carried out. The objectives of the experiment are to find out which trained model gives the best performance in terms of accuracy and losses. The accuracy is the number of correctly predicted by the model. The accuracy rate was evaluated by using the confusion matrix. While the loss rate is cumulative of the mean square loss of the dataset.

Two sample sizes were selected (320x240 and 80x60) with identical dataset split proportion (training: 70%; validation: 10%; test: 20%). Every model will be trained in epochs and three models had been built for each sample size category. The best model selection criteria for each model were depending on its best-fitting epoch from 1-30 epoch(s). The evaluation performance of each model had been carried out as tabulated separately in Tables 1-2 by sample frame size categories.

Table 1. Model performance for 320x240 frame size

Model	Performance (%)			
	Validation		Test	
	Accuracy	Validation loss	Accuracy	Test loss
A.1	92.98	16.47	87.61	35.54
A.2	87.72	13.76	88.89	17.02
A.3	*98.18*	*41.6*	*96.43*	*10.3*

Table 2. Model performance for 80x60 frame size

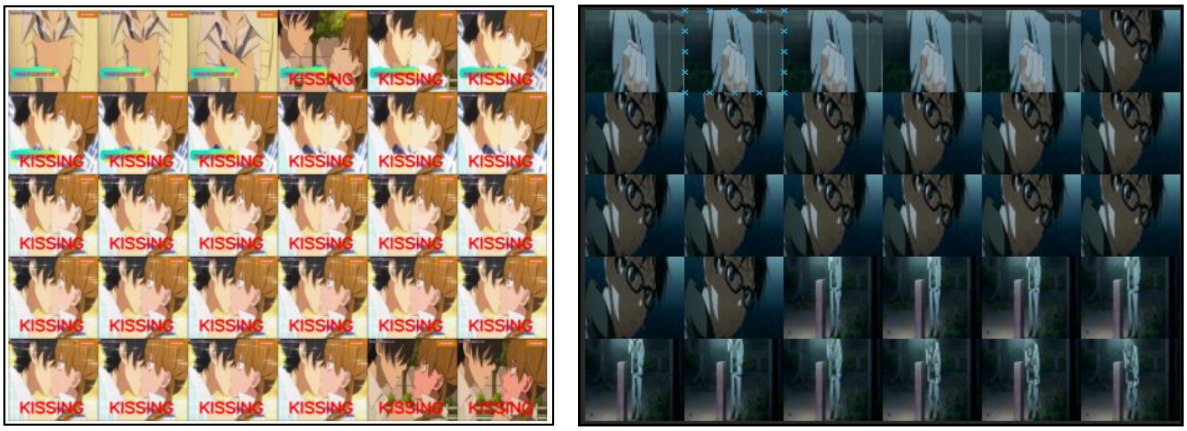
Model	Performance (%)			
	Validation		Test	
	Accuracy	Validation loss	Accuracy	Test loss
B.1	83.33	32.49	73.91	50.12
B.2	92.59	11.56	85.96	18.05
B.3	*98.28*	*4.84*	*92.24*	*11.03*

Based on the performance rate in Table 1 (Model-320x240), Model A.3 has the highest validation and test accuracy rate with 98.18 % and 96.43%; respectively. While the performance rates for Model A.1 and Model A.2 are 92.98%; 87.61% and 87.72%; 88.89%; respectively. Furthermore, Model A.1 also has the lowest mean test loss with 10.3%; compared to Model A.1 (17.02%) and A.2 (17.02%). Thus, Model A.3 is the best model amongst the Model-320x240 category. Whereas for Model-80x60 (Table 2), the result shows that Model B.3 gave the best performance in validation and test accuracy (98.28%; 92.24%) with a mean loss rate of 4.84% and 11.03%; respectively. While Model B.1 and B.2 performed well in an acceptable accuracy rate (>73%) and mean loss rate (<50%).

The fitness value is the difference between train loss and validation loss. Table 3 shows the performance of the best two models (A.3 and B.3). The fitness value for Model B.3 (2.29%) is less than Model A.3 (2.68%). This means the performance of Model B.3 is better than Model A.3 in terms of robustness and fitness. However, the difference in fitness rates for both models (0.39%) was not too significant to assume that Model B.3 is performed well than A.3. Moreover, the fitness rate of Model A.3 is lower than Model B.3; which justifies the lower the mean loss of the model, the closer the predictions to the true labels. Therefore, Model A.3 is considered the best model in terms of high accuracy and fitness performance amongst the six models. The deployment of the model is executed to detect the kissing and non-kissing scenes in video clips. Figure 5 shows the model output result samples. Figure 5(a) shows sample from *tonari no kaibatsu kun* and Figure 5(b) shows sample from *shigatsu wa kimi no uso*.

Table 3. Comparison of best model performance

Model	Frame Size	Loss rate (%)		Fitness
		Training	Validation	value (%)
A.3	320x240	1.48	4.16	2.68
B.3	80x60	2.55	4.84	2.29



(a) (b)

Figure 5. Samples of A.3 model output results (a) '*tonari no kaibatsu kun'* video clip and (b) '*shigatsu wa kimi no uso*' video clip

To avoid bias, the test video clips were selected from other than the clips in the dataset. The test video clip parameters are 320x240 pixels resolution, 30 frames per second, and the duration range of 3-12 seconds. The deployment shows that the system capable to distinguish the difference between kissing and non-kissing scenes. However, the model has misclassified some early frames of the kissing scene, where it appears kissing-like behavior features or non-kissing clips that stand close between each other in the visual as shown in Figure 5 (a): Frame-4.

4. CONCLUSION

A cartoon can be much more perilous than any other experience for twelve years old children. It may contain content that confuses the child which contains directions that may contradict the religion and culture. This may lead the children to have a different undesirable point of view of her/his parents, teachers, and even his lord. With the state of technology growth, this project could be one of the features in parenting software to filter out unnecessary explicit scenes such as kissing. In this study, the combination pre-trained model of CNN Darknet-19 and conv-LSTM can perform well in detecting the explicit kissing scene in a cartoon video. The model is able to extract the spatial-temporal features that enable the analysis of the local motion of the character taking place in the video. The final trained model with 320x240 pixels image was able to achieve 96.43% of accuracy. Therefore, this kissing scene detector can be used as a filter for digital video player applications. For continuous improvement throughout this project, we are proposing to build more explicit scene filters particularly for a digital video/ photo that may contain pornography, violence, and abuse features; which research findings have proven to have a destructive effect on a child.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support from Short Term Research Grant (Str19023) UniKL.

REFERENCES

[1] S. Ahmed and J. A. Wahab, “Animation and socialization process: Gender role portrayal on cartoon network,” *Asian Social Science*, vol. 10, no. 3, pp. 44-53, 2014, doi: 10.5539/ass.v10n3p44.

[2] K. Habib and T. Soliman, “Cartoons’ Effect in Changing Children Mental Response and Behavior,” *Open Journal of Social Sciences*, vol. 03, no. 9, pp. 248-264, 2015, doi: 10.4236/jss.2015.39033.

[3] Malaysian Communications and Multimedia Commission, “The Malaysian Communication and Multimedia Content Code,” 2004.

[4] T. Hughes, “Unrestricted Film List Project,” Hispania, 2007, doi: 10.2307/20063560.

[5] Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” 14th *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078468.

[6] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, “Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition,” *Pattern Recognitition*, vol. 76, pp. 80-94, 2018, doi: 10.1016/j.patcog.2017.10.033.

[7] G. W. Nie, N. F. Ghazali, N. Shahar, and M. A. As’Ari, “Deep stair walking detection using wearable inertial sensor via long short-term memory network,” *Bulletin of Electrical Engineering and Informatic*, vol. 9, no. 1, pp. 238-246, 2020, doi: 10.11591/eei.v9i1.1685.

[8] A. Ziai, “Detecting Kissing Scenes in a Database of Hollywood Films,” *Computer Science > Computer Vision and Pattern Recognition*, 2019, [Online]. Available: http://arxiv.org/abs/1906.01843.

[9] I. A. M. Zin, Z. Ibrahim, D. Isa, S. Aliman, N. Sabri, and N. N. A. Mangshor, “Herbal plant recognition using deep convolutional neural network,” *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 2198-2205, 2020, doi: 10.11591/eei.v9i5.2250.

[10] Y. Pratama, E. Marbun, Y. Parapat, and A. Manullang, “Deep convolutional neural network for hand sign language recognition using model E,” *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 5, pp. 1873-1881, 2020, doi: 10.11591/eei.v9i5.2027.

[11] W. Setiawan, M. I. Utoyo, and R. Rulaningtyas, “Reconfiguration layers of convolutional neural network for fundus patches classification,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 1, pp. 383-389, 2021, doi: 10.11591/eei.v10i1.1974.

[12] M. A. Rasyidi, R. Handayani, and F. Aziz, “Identification of batik making method from images using convolutional neural network with limited amount of data,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 1300-1307, 2021, doi: 10.11591/eei.v10i3.3035.

[13] A. H. Abdulnabi, G. Wang, J. Lu and K. Jia, “Multi-Task CNN Model for Attribute Prediction,” *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949-1959, 2015, doi: 10.1109/TMM.2015.2477680.

[14] E. B. Nievas, O. Deniz Suarez, G. Bueno García, and R. Sukthankar, “Violence Detection in Video Using Computer Vision Techniques,” *International Conference on Computer Analysis of Images and Patterns*, 2011.

[15] K. Charalampous and A. Gasteratos, “On-line deep learning method for action recognition,” *Pattern Analysis and Applications*, vol. 19, no. 2, pp. 337-354, 2016, doi: 10.1007/s10044-014-0404-8.

[16] W. Sultani, C. Chen, and M. Shah, “Real-World Anomaly Detection in Surveillance Videos,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479-6488, doi: 10.1109/CVPR.2018.00678.

[17] M.F. Abu Hassan, A. Hussain, and M.H. Md Saad, *Fitur Poligon: Pengawasan Aksi Insan berasaskan Video Pintar*. Penerbit UKM, 2021. Accessed on: Jan. 4, 2022. [Online]. Available: https://ukmpress.ukm.my/index.php?route=product/product&path=90&product_id=1374book

[18] A. G. Mahmoud, A. M. Hasan, and N. M. Hassan, “Convolutional neural networks framework for human hand gesture recognition,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2223-2230, 2021, doi: 10.11591/eei.v10i4.2926.

[19] M. A. Obaid and W. M. Jasim, “Pre-convoluted neural networks for fashion classification,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 750-758, 2021, doi: 10.11591/eei.v10i2.2750.

[20] L. OpenShot Studios, “OpenShot Video Editor.” https://www.openshot.org/ (accessed Aug. 15, 2021).

[21] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” 2016, [Online]. Available: http://arxiv.org/abs/1612.08242.

[22] T. Emara, H. E. A. El Munim, and H. M. Abbas, “LiteSeg: A Novel Lightweight ConvNet for Semantic Segmentation,” *Digital Image Computing: Techniques and Applications (DICTA)*, 2019, doi: 10.1109/DICTA47822.2019.8945975.

[23] P. Zhu, H. Wang and V. Saligrama, “Zero Shot Detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 998-1010, 2020, doi: 10.1109/TCSVT.2019.2899569.

[24] Z. Dong, J. Qin, and Y. Wang, “Multi-stream Deep Networks for Person to Person Violence Detection in Videos,” *Chinese Conference on Pattern Recognition*, 2016, vol. 662, doi: 10.1007/978-981-10-3002-4.

[25] M. F. Abu Hassan, A. Hussain, M. H. Muhamad, and Y. Yusof, “Convolution neural network-based action recognition for fall event detection,” *International Journal Advanced Trends Computer Science Engineering*, vol. 8, no. 1, pp. 466-470, 2019, doi: 10.30534/ijatcse/2019/6881.62019.

BIOGRAPHIES OF AUTHORS

Muhammad Arif Haikal Muhammad Fadzli is a final year Bachelor of Engineering Technology (HONS) in Mechatronics study in Universiti Kuala Lumpur Malaysia France Institute (UniKL MFI). His research interest is in image processing and machine learning for software development. He can be contacted at email: haikalfadzli2305@gmail.com.

Mohd Fadzil Abu Hassan is a senior lecturer at Universiti Kuala Lumpur. Awarded Doctor of Philosophy degree from Universiti Kebangsaan Malaysia in 2019. The academic and research mastery area is primarily focusing on machine vision and artificial intelligence for smart surveillance and monitoring systems. Obtained funds and support from local universities and industry for innovative engineering research and system development. He can be contacted at email: fadzil@unikl.edu.my.

Norazlin Ibrahim primary field of research interest is Artificial Intelligence (AI). Currently, she is a senior lecturer at Universiti Kuala Lumpur and holds a Doctor of Philosophy degree from Universiti Kebangsaan Malaysia. Within AI, she is interested in problems related to vision, machine learning, and data mining towards interdisciplinary applications such as biomedical and agriculture sciences. Holding funds from local universities for engineering research and system development. She can be contacted at email: norazlin@unikl.edu.my.