The Best Model of Convolutional Neural Networks Combined with LSTM for the Detection of Interpersonal Physical Violence in Videos

Hugo Calderon-Vilca
Universidad Nacional Mayor de San Marcos
Lima, Perú
hcalderonv@unmsm.edu.pe

Kent Cuadros Ramos
Universidad Nacional Mayor de San Marcos
Lima, Perú
kent.cuadros@unmsm.edu.pe

Elmer Diaz Quiroz
Universidad Nacional Mayor de San Marcos
Lima, Perú
elmer.diaz2@unmsm.edu.pe

Jorge Angeles Rojas
Universidad Nacional Mayor de San Marcos
Lima, Perú
jorge.angeles4@unmsm.edu.pe

René Calderon Vilca
Universidad Nacional de San Agustín
Arequipa, Perú
acalderon@defondo.com

Alejandro Apaza Tarqui
Universidad Nacional del Altiplano
Puno, Perú
apazatarqui@unap.edu.pe

***Abstract*—Citizen insecurity is directly related to interpersonal physical violence, there are algorithms that allow detecting violence in videos; therefore, it is necessary to know which is the best model for detecting violence. For this research, three convolutional neural network models were compared: Xception, InceptionV3 and VGG16 each together with a recurrent LSTM network, to find out which of the models is the best for the detection of interpersonal violence in videos. The three models were trained using the Real Life Violence Situations dataset, then violence and non-violence were classified, as a result, the InceptionV3 model is the best model, managing to classify with an accuracy of 94% compared to the VGG16 and Xception models, which obtained 88% and 93% respectively. Therefore, we recommend the InceptionV3 model for the detection of interpersonal physical violence in citizen security videos.**

I. Introduction

Citizen insecurity is one of the main problems in Peru as in all of Latin America, in addition to poverty and economic growth. Citizen insecurity is commonly related to interpersonal physical violence [1]. As a preventive measure due to the increase in citizen insecurity, society has been seeking solutions to its security systems, considering the monitoring of human actions that are related to physical violence. This monitoring is carried out through video surveillance systems which are supervised by security entities or ordinary people.

Currently, artificial intelligence algorithms allow people to solve different problems, such as heart disease detection with neural networks [14], tuberculosis detection with image processing [18], information search and retrieval [15], augmented reality for museums [16], routing questions and answers [17], recommendation of videogames with fuzzy logic [19] and many other purposes.

Video surveillance systems are frequently installed in urban areas to support the security and surveillance of citizens. In the cameras of these video surveillance systems, it is possible to record scenes of physical violence and the scenes are observed and analyzed by the personnel in charge to detect the existence of violence according to their judgment.

In [5] the development of an intelligent system of recognition of human activity is proposed by building a robust neural network based on the database of human activities, in [4] a pedestrian detection system based on HOGG (Gabor Filters and Histogram of Oriented Gradient (HOG)) for feature extraction is proposed, obtaining better results than only HOG or only the Gabor filter and finally the Convolutional Neural Networks (CNN) for pedestrian detection, in [3] a method for the system of automatic recognition of human activity through walking (identifying an activity by the way they walk) without human intervention is proposed, this system is based on foreground extractions, people tracking, feature extraction and recognition.

From the investigations [5], [4], [3] convolutional networks are applied for the preprocessing of images and extraction of significant features and, therefore, they do not apply for the detection of violence.

For the purpose of detecting interpersonal physical violence in citizen security videos, in research [10], [11], [2] the Hockey dataset is used, which is not directly related to situations of violence in real life, but to situations of violence in hockey games.

The VGG [22] model for facial recognition in real-world surveillance videos, allows face detection. The VGG model comprises eight convolutional layers and three fully connected layers, it uses ReLU and max-pooling for its classification.

ISSN 2305-7254

The Inception model proposed in [21] a deep convolutional neural network structural to classify movements.

They also managed to automatically detect dangerous situations to guarantee the safety of residents in the surveillance areas [25], behaviors such as vandalism, brawl, robbery and others are considered.

Another Xception model proposed by [20] a deep learning architecture with separable convolutions in depth inspired by the Inception model, slightly exceeds Inception V3, this model has been tested with 350 million images and with 17,000 classes.

The VGG16 model proposed by [23] is a CNN that achieves 92.7% accuracy in ImageNet, tested with 14 million images with 1000 classes.

In research [12] the new improved algorithm Inception V3 is presented, a CNN based on the home network that allows detecting common characteristics of both dark and light images, 48 thousand images of ships in 9 categories were tested, achieving an improvement of 17.48% to the original Inception algorithm.

The Xception [20] and VGG16 [23] models were tested with millions of images, their classes were 17000 and 1000 respectively, as for Inception V3 [12] 48000 images with 9 classes were tested. On the other hand, the detection of interpersonal violence has several characteristics that could be associated with these 3 models, so it is necessary to test which of the models would be the best to detect interpersonal violence in the videos.

In this research, our aim is to compare three models of Convolutional Neural Networks: Xception, VGG16 and InceptionV3 to determine the best artificial vision model with the objective of detecting interpersonal physical violence in citizen security videos. In each artificial vision model, the image sequences were processed using the videos of the Kaggle Real Life Violence Situations Dataset, image processing with the extraction of significant features through a convolutional neural network was performed, then these pre-processed features passed to the LSTM network sequentially for the analysis of the existence of violence and, finally, they were classified.

This paper is organized as follows: Section II of this document deals with related works as a State of the Art are presented in Section. Then, a comparison of the three models: Xception, VGG16 and InceptionV3 is presented in Section III. Our experimental evaluation and results are described in Section IV and, finally, the conclusion and future work are presented in the last part of the document.

II. State of the Art

Currently in the surveillance systems, various solutions supported by the different Information Technologies are being applied to obtain better results in their monitoring, seeking the automation of tasks such as the detection and recognition of human activities that are mainly related to situations of citizen insecurity with the presence of violence and that are totally independent of any human intervention. For a better understanding of the technologies that could help us address this problem, a specialized search of scientific articles was carried out, selecting papers to validate and solve the problem, all of them focused on the field of Machine Vision, thus allowing us to review the techniques that are currently used in this field. The selected articles provide information in different scenarios.

A. Models for event detection

In research [6] the detection of events was carried out by means of the video summary, the video summary is a function of human properties, not linked to violence. In research [7], [2] for the detection of events related to violence in real time, a convolutional neural network (CNN) was used, [7] in addition to estimating the perceived violence, image feelings in the emotional dimensions were analyzed, [6], [2] analyzed the video image sequences and in [7] they worked with static images.

B. Applications in surveillance systems

In research [8], [4], [5], neural networks were used for the application of the surveillance systems. In [4] they focused on the detection of pedestrians, in [5] they managed to detect and track the human body recognizing the type of movement and daily human activity (shaking hands, lying, walking, sitting, boxing) and [8] worked on the detection of moving objects. [8], [4], [5] carried out image preprocessing as a previous step, [8] in addition to performing image processing techniques, a new unified technique (the region of interest (ROI) finder and YOLO which is a depth detector) was applied, further improving their image sequence processing speed. [10], [9] focused on detecting violence in surveillance systems, in [10] it is done using a semi-supervised learning framework to classify whether a behavior is violent or not, in [9] it is done through the audio of the recordings from the different data sources worked on the violence.

III. Methodology to compare the convolutional network models: Inception v3, vgg16 and xception

In this section, the components of the Convolutional Neural Network models for detecting interpersonal physical violence in citizen security videos are presented.

A. Preparation of the Dataset

The dataset contains a set of videos of situations of interpersonal physical violence in real life, with a variety of people of different race, age and gender, and also has variety in the environment. The dataset was downloaded from KAGGLE, which is an online community of data scientists and machine learners, owned by Google LLC. The Real Life Violence Situations dataset contains 1000 small videos of violence and 1000 small videos of non-violence with a maximum duration of 7 seconds and an average duration of 5 seconds.

B. Video pre-processing

- The video was divided into image sequence: from the dataset, each video into 20 image sequences.
- The extracted image sequences were resized, for the Xception and Inceptionv3 convolutional neural networks the boxes were resized to (299 * 299 * 3) and for VGG16 to

(224 * 224 * 3), since each network requires a certain type of image size respectively.

- Shuffle Data: the data was shuffled to prevent the model from biased learning of a certain data pattern.

C. Feature extraction

20 sequences of images extracted from the video are entered and processed in batches, passing through the different layers of the convolutional network model being used:

- Xception: it goes from the first layer "input_1" to the pooling layer "avg_pool", obtaining the significant features of the intermediate layer "avg_pool" of each image sequence.
- InceptionV3: it goes from the first layer "input_1" to the pooling layer "avg_pool", obtaining the significant features of the intermediate layer "avg_pool" of each image sequence.
- VGG16: it goes from the first layer "input_1" to the "flatten" layer, obtaining the significant features of the intermediate layer "flatten" of each image sequence.

These significant features are considered as transfer values that are used as the input to the LSTM Neural Network. From each image sequence, a vector of transfer values is obtained as an output depending on the convolutional network model, for Xception, InceptionV3 and VGG16 a vector with 2048, 2048, 25088 transfer values respectively was obtained, as from each video 20 image sequences were processed, so (*20 x n*) video transfer values were obtained, where *n* is the vector with the transfer values.

D. Classification with LSTM and fully connected layers

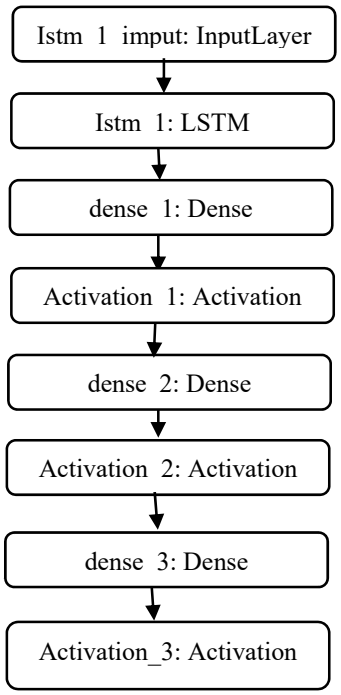

Fig. 1. Applying the recurrent neural network LSTM

For the classification, the second neural network was trained using the classes of the violence dataset (Violence, Non-Violence), so that the neural network learned to classify images based on the transfer values of the convolutional neural network model that was being used.

The LSTM component has a size of 512 neurons used to extract temporal features, the features are extracted throughout the sequence of images obtained from the videos of surveillance systems. Fig. 1 illustrates the details of the LSTM architecture.

The input form of the LSTM is (None, 20, n), 20 represents the number of image sequences extracted from the videos of the Real Life Violence Situations dataset and *n* is the size of the vector with the transfer values obtained from the convolutional network used.

The classification was carried out considering the 20 image sequences obtained from the video. If any of them presented violence, the video was classified as violent.

E. Design of the architecture of the convolutional network model with LSTM

Architecture to determine the best convolutional neural network model that allows the detection of interpersonal physical violence in citizen security videos is made up of three parts, preprocessing, feature extraction and classification.

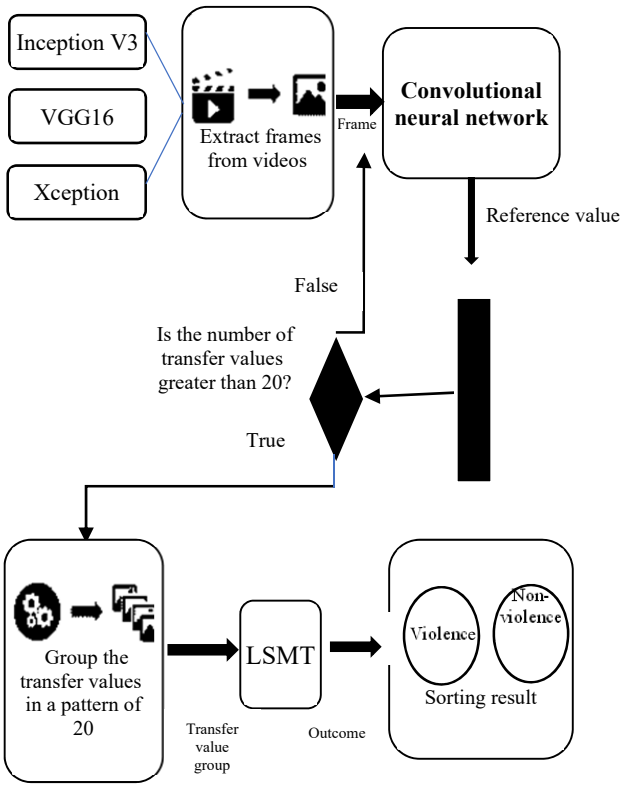

Fig. 2. Artificial Vision Model to classify violence

In the pre-processing, sequences of images from a video file were obtained converting them to a suitable format for the neural network, this was done in order for the network to train

efficiently. For feature extraction, the image sequences passed through the convolutional neural network. Then, for the classification, the significant features obtained from the image sequences passed through the LSTM Recurrent Neural Network for the extraction of temporal features, followed by two hidden layers and the output layer is a layer of 2 neurons with softmax activation, which gives us the final classification (the presence of violence or not).

IV. Results and Discussion

This section describes the implementation of a dataset as well as the results and comparison of the convolutional neural network models: Xception, VGG16 and InceptionV3.

For the experiment, The Real Life Violence Situations dataset was used with 2000 videos (1000 with violence 1000 non-violence), of which 1600 were divided for the training of the models and 400 for the test. No data augmentation techniques were performed, only the original videos are used.

We carried out the training with models Xception, InceptionV3 and VGG16, combining each model with LSTM to classify the existence or non-existence of violence according to the architecture proposed in Fig. 2.

In the following graphs the accuracy and loss in the training process of the model by epochs is shown respectively. Then, an evaluation of which of the models had better accuracy and less loss was performed.

The results of the experiment entering the the Xception model are shown, in Fig. 3 you can see the behavior of the accuracy metric and in Fig. 4 the loss in training.

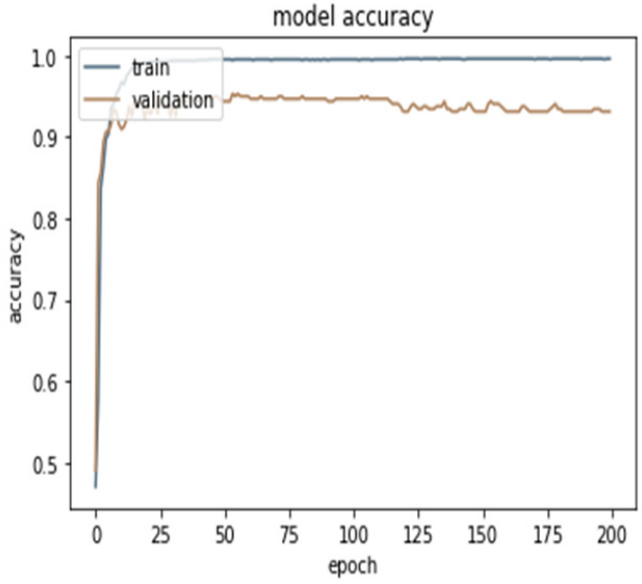

Fig. 3. Xception model accuracy

In 200 epochs, from the graph of the Xception model, an average accuracy of approximately 93% is achieved.

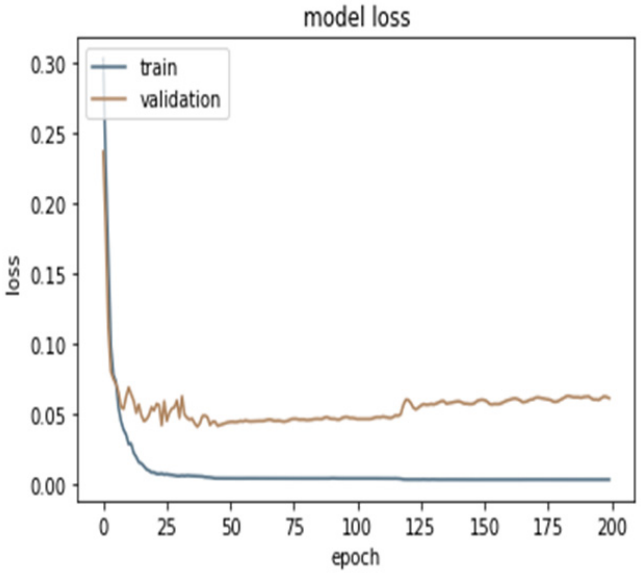

Fig. 4. Xception model loss

Regarding the loss for the Xception model, there is 7% during training.

The results of the experiment entering the Inception V3 model are shown, in Fig. 5 you can see the behavior of the accuracy metric and in Fig. 6 the loss in training.

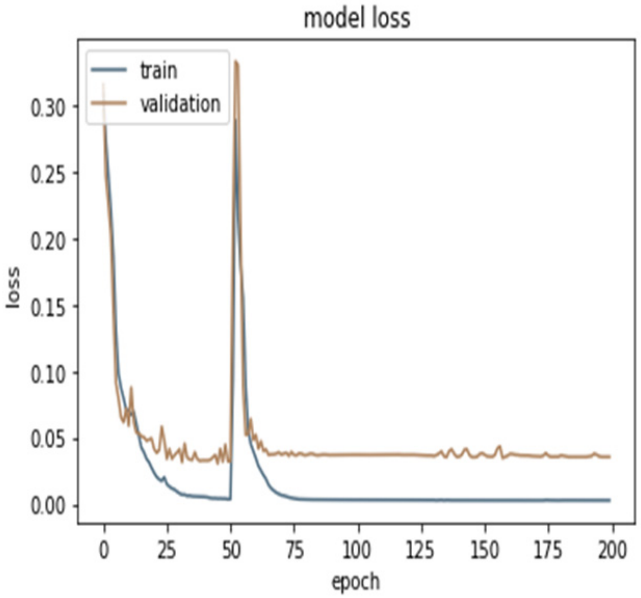

Fig. 5. InceptionV3 model accuracy

In 200 epochs, from the graph of the Inception V3 model, an average accuracy of approximately 94% is achieved.

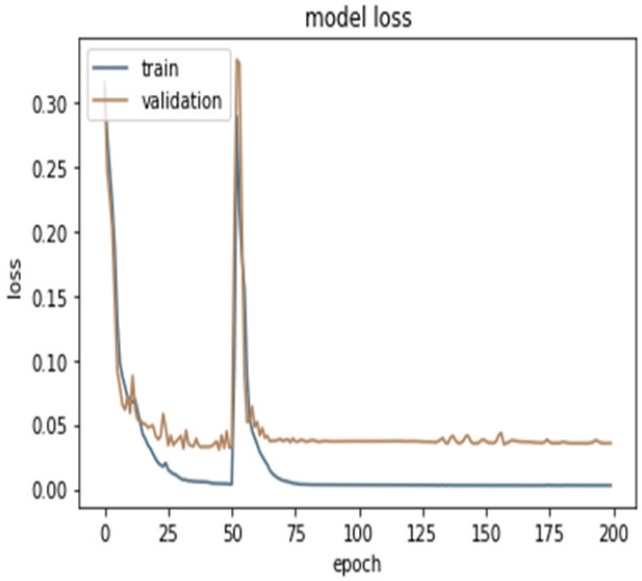

Fig. 6. InceptionV3 model loss

Regarding the loss for the Inception V3 model, there is 5% during training.

The results of the experiment entering the VGG16 model are shown, in Fig. 7 you can see the behavior of the accuracy metric and in Fig. 8 the loss in training.

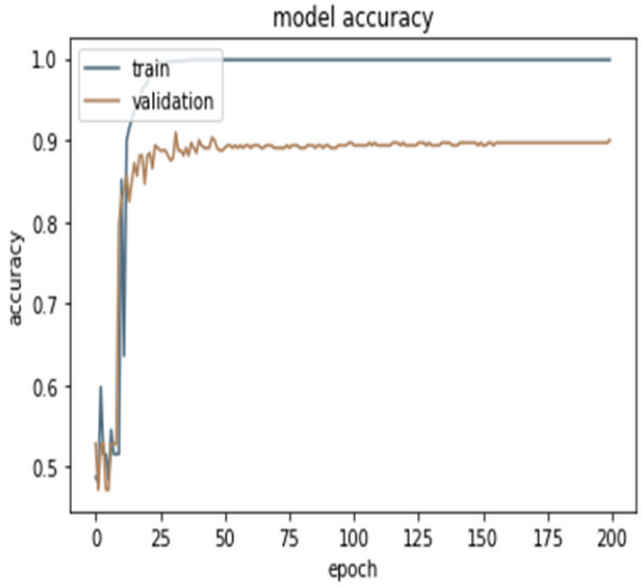

Fig. 7. VGG16 model accuracy

In 200 epochs, from the graph of the VGG16 model, an average accuracy of approximately 88% is achieved.

Regarding the loss for the VGG16 model, there is 10% during training.

After training, the test data was submitted, then the results were shown. The confusion matrix of each of the models in Table I was also evaluated.

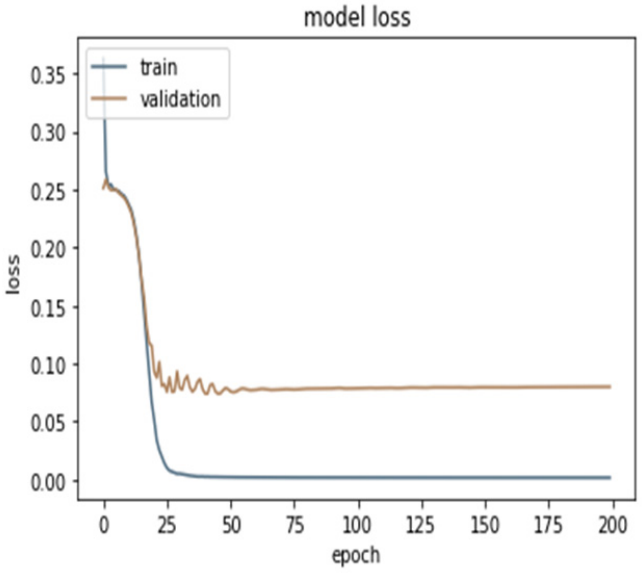

Fig. 8. VGG16 model loss

TABLE I. XCEPTION CONFUSION MATRIX

	Non-violence	Violence
Non-violence	197	8
Violence	20	175

TABLE II. INCEPTIONV3 CONFUSION MATRIX

	Non-violence	Violence
Non-violence	205	11
Violence	13	171

TABLE III. VGG16 CONFUSION MATRIX

	Non-violence	Violence
Non-violence	197	32
Violence	15	156

As it can be seen from the previously presented models: Xception, InceptionV3 and VGG16 presented an average accuracy 93%, 94% and 88% respectively. In addition, it was observed that the InceptionV3 model obtained greater sensitivity in detecting violence. Analyzing the results of the experiment it was determined that the InceptionV3 model had better results compared to Xception and VGG16. Therefore, the best model to detect interpersonal physical violence in videos is the InceptionV3 model, combined with LSTM, which classifies violence or non-violence in the videos, it is highly recommended that researchers use this model to solve these types of problems.

In research [12] an improved InceptionV3 for classification of obscured ships in remote sensing images is presented, reaching 70% accuracy. Comparing our proposed model InceptionV3 for the detection of violence in videos, 94% accuracy is a great improvement. Also, in the work of Intelligent Recognition of Fatigue and Sleepiness Based on InceptionV3-LSTM [13] they show 87.5% accuracy. Our research reached 94% accuracy, which is also a great improvement.

At research [7] Protest Activity Detection and Perceived Violence Estimation from Social Media Images is proposed. If we compare, the best model for the detection of interpersonal physical violence in videos is the one proposed in our research.

In research [8], [4], [5] they used neural networks for the application of surveillance systems, in our research the best model for the detection of interpersonal violence in videos was determined, our investigation could be used for the implementation of the validity systems. In the proposal for the detection of violence in surveillance systems [10] it is done through a semi-supervised learning framework to classify whether a behavior is violent or not, in our research, the existence of physical interpersonal violence in the videos is classified.

CONCLUSION AND FUTURE WORK

In this research, three models of convolutional neural networks were analyzed, and it was determined which was the best model for detecting interpersonal physical violence in citizen security videos. When comparing the three models, it could be seen that the InceptionV3 model performs better for the detection of violence in the citizen security videos. In the future, work could be done on an integrated application for municipalities that can give notice or alert authorities to the existence or not of interpersonal violence in the streets. This application would be integrated with the best proven InceptionV3 model.

REFERENCES

[1] Carlos Reyna y Eduardo Toche (1999). Comisión Económica para América Latina y el Caribe, la inseguridad en el Perú. serie políticas sociales. https://core.ac.uk/download/pdf/45620332.pdf

[2] Sudhakaran, S., & Lanz, O. (2017). Learning to detect violent videos using convolutional long short-term memory. 2017 14Th IEEE International Conference On Advanced Video And Signal Based Surveillance (AVSS). doi: 10.1109/avss.2017.8078468.

[3] Prakash Gupta, J., Dixit, P., & Bhaskar Semwal, V. (2014). Analysis of Gait Pattern to Recognize the Human Activities. International Journal Of Interactive Multimedia An Artificial Intelligence, 2(7), 7. doi: 10.9781/ijimai.2014.271.

[4] Ahmed, F., Topu, B., & Islam, S. (2019). HOG and Gabor Filter Based Pedestrian Detection using Convolutional Neural Networks. 2019 International Conference On Electrical, Computer And Communication Engineering (ECCE). doi: 10.1109/ecace.2019.8679133.

[5] Babiker, M., Khalifa, O., Htike, K., Hassan, A., & Zaharadeen, M. (2017). Automated daily human activity recognition for video surveillance using neural network. 2017 IEEE 4Th International Conference On Smart Instrumentation, Measurement And Application (ICSIMA). doi: 10.1109/icsima.2017.8312024.

[6] Thomas, S., Gupta, S., & Subramanian, V. (2017). Smart surveillance based on video summarization. 2017 IEEE Region 10 Symposium (TENSYMP). doi: 10.1109/tenconspring.2017.8070003.

[7] Won, D., Steinert-Threlkeld, Z., & Joo, J. (2017). Protest Activity Detection and Perceived Violence Estimation from Social Media Images. Proceedings Of The 2017 ACM On Multimedia Conference - MM '17. doi: 10.1145/3123266.3123282.

[8] Muchtar, K., Rahman, F., Munggaran, M., Dwiyantoro, A., Dharmadi, R., & Nugraha, I. (2019). A unified smart surveillance system incorporating adaptive foreground extraction and deep learning-based classification. 2019 International Conference On Artificial Intelligence In Information And Communication (ICAIIC). doi: 10.1109/icaiic.2019.8669017

[9] Roa, J., Jacob, G., Gallino, L., Hung, P.C.K. Towards Smart Citizen Security Based on Speech Recognition. 2018 Congreso Argentino de Ciencias de la Informática y Desarrollos de Investigación (CACIDI), Buenos Aires, 2018, pp. 1-6.

[10] Zhang, T., Jia, W., Gong, C., Sun, J. and Song, X. (2018). Semi-supervised dictionary learning via local sparse constraints for violence detection. Pattern Recognition Letters, 107, pp.98-104

[11] Fillipe D. M. de Souza , Guillermo C. Chavez , Eduardo A. do Valle Jr. , Arnaldo de A. Araujo, Violence Detection in Video Using Spatio-Temporal Features, Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images, p.224-230, August 30-September 03, 2010 doi:10.1109/SIBGRAPI.2010.38.

[12] K. Liu, S. Yu and S. Liu, "An Improved InceptionV3 Network for Obscured Ship Classification in Remote Sensing Images," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 4738-4747, 2020, doi: 10.1109/JSTARS.2020.3017676.

[13] Y. Zhao, K. Xie, Z. Zou and J. -B. He, "Intelligent Recognition of Fatigue and Sleepiness Based on InceptionV3-LSTM via Multi-Feature Fusion," in IEEE Access, vol. 8, pp. 144205-144217, 2020, doi: 10.1109/ACCESS.2020.3014508.

[14] H. D. Calderon-Vilca, K. E. C. Callupe, R. J. I. Aliaga, J. B. Cuba and F. C. Mariño-Cárdenas, "Early Cardiac Disease Detection Using Neural Networks," 2019 7th International Engineering, Sciences and Technology Conference (IESTEC), Panama, 2019, pp. 562-567.

[15] M. A. Silva-Fuentes, H. D. Calderon-Vilca, E. F. Calderon-Vilca and F. C. Cárdenas-Mariño, "Semantic Search System using Word Embeddings for query expansion," 2019 IEEE PES Innovative Smart Grid Technologies Conference - Latin America (ISGT Latin America), Gramado, Brazil, 2019, pp. 1-6.

[16] H. D. Calderon-Vilca, E. G. Ocrospoma-Callupe, F. C. Cárdenas-Mariño and E. F. Calderon-Vilca, "Architecture and mobile application with augmented reality, visualizing videos and 3d objects in museums," 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), Valparaiso, Chile, 2019, pp. 1-5.

[17] C. E. Serquen-Llallire, H. D. Calderon-Vilca and F. C. Cardenas-Mariño, "Comparison of two Algorithms for Routing Questions and Answers, Applied to Group of Students Software Engineering," 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Gudalajara, Mexico, 2018, pp. 1-6.

[18] H. D. Calderon-Vilca, Luis M. Ortega Melgarejo and F. C. Cardenas-Mariño, "Tuberculosis Detection Architecture with Image Processing using the SIFT and K-Means Algorithm", Journal Computación y Sistemas, Vol. 24, N° 3, 2020, pp. 989-997.

[19] H. Calderon-Vilca, N. M. Chavez and J. M. R. Guimarey, "Recommendation of Videogames with Fuzzy Logic," 2020 27th Conference of Open Innovations Association (FRUCT), Trento, Italy, 2020, pp. 27-37, doi: 10.23919/FRUCT49677.2020.9211082.

[20] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-January, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[22] Ya, W., Bao, T., Chunhui, D., & Zhu, M. (2017). Face recognition in real-world surveillance videos with deep learning method. 2017 2nd International Conference on Image, Vision and Computing, ICIVC 2017, 239–243. https://doi.org/10.1109/ICIVC.2017.7984553

[23] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. Paper presented at the 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings

[24] Ryabchikov, I., & Teslya, N. (2021). Estimating position of multiple people in common 3d space via city surveillance cameras. Paper presented at the Conference of Open Innovation Association, FRUCT, , 2021-January doi:10.23919/FRUCT50888.2021.9347579