ISSN 2278-3091
Volume 8, No.1.6, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse6881.62019.pdf
https://doi.org/10.30534/ijatcse/2019/6881.62019

Convolution Neural Network-based Action Recognition for Fall Event Detection

Mohd Fadzil Abu Hassan[1], Aini Hussain[2], Muhamad Hanif Md. Saad[3], Yusman Yusof[4]
[1,4]Industrial Automation Section, Universiti Kuala Lumpur Malaysia France Institute, Malaysia
{fadzil, yusman}@unikl.edu.my
[2,3]Centre for Integrated Systems Engineering and Advanced Technologies, Universiti Kebangsaan Malaysia, Malaysia
{draini, hanifsaad}@ukm.edu.my

ABSTRACT

Action recognition is a challenging and essential task in computer vision particularly in monitoring elderly activities in daily living. This paper presents an efficient normal and abnormal actions recognition model, particularly for fall event detection using the convolution neural network; CNN-AlexNet. In this study, the model was trained and tested using an established CASIA and Le2i datasets that cover various application domains. Results showed that the model performed well with classification accuracy, F-score, sensitivity and specificity rates of 99.97%, 99.97%, 99.94% and 100%, respectively. Therefore, the CNN-AlexNet based action recognition model can be considered as a high-performance classifier to differentiate between normal and abnormal actions to be adapted in a smart camera-based surveillance system.

Key words: Action recognition, CNN-AlexNet, convolution neural network, fall event.

1. INTRODUCTION

Over the years, healthcare and socioeconomic in Malaysia have seen tremendous improvements in its healthcare delivery system. Thus, it contributes to the rise in population aging (also known as 'Silver Tsunami'), which is showing a positive indicator of decline in early age mortality. According to [1], life expectancy at this age had risen due to advancements in public health and medical technologies, along with the improvements in quality of life. Due to these positives factors, the population rate is projected to continue growing over 15% over the years and Malaysia is expected to become an aging nation by 2030 [2]. The projected Malaysian population shows the number of seniors (>60 years) will be equal to the number of young people (<15 years) by 2045, whereby each group representing 20% of the total population in Malaysia. The proportion of seniors is expected to continue growing sharply by 49% with a forecast of 9.6 million by 2050 as compared to 4.9 million in the previous 20 years (2030).

According to [3], 9.0% of senior citizens live alone and 20.9% live together with his/her spouse. While the majority of seniors (70%) are supervised by their family members and senior care centers. Most senior citizens have health problems involving non-communicable diseases (NCD) such as heart disease, osteoarthritis, stroke, diabetes, Parkinson's etc. [2]. These health problem factors faced by seniors tend to cause harm in their daily activities such as exposure to the risk of falls which may lead to bone fractures. As stated in [4], the mortality rates of senior citizens after attending the emergency department due to fall accidents within 1, 2, 5, and 10 years were 22%, 37%, 49%, and 80%, respectively. From these findings, the mortality rate has continuously increased over the last 10 years and most of the fall events happen at home; although this area is considered a safe and secure. Therefore, precautions and rapid action should be taken by the caregivers to minimize the risk of further injury due to falls. Thus, it may reduce morbidity risk, treatment costs as well as mortality rate.

Rapid growth of machine vision technology has greatly influenced many areas of research, especially computer technology and software. This improvement gives a positive impact on human life i.e. the use of semi-automated video surveillance systems (VSS) in monitoring human activities. However, performing the manual task of real-time VSS requires a high degree of visual focus, as well as a high cost of remuneration. Computer vision-based system has been proven to help provide useful information to automatically monitor human movements; not only in public areas and workplaces but in a safe place within residential areas. However, most VSS for home security and small premises are not fully automated because the monitoring and evaluation of the activities that take place are closely monitored by the guardian or human operator [5].

Convolution neural network (CNN) is an artificial neural network-based machine learning algorithm [6] and one of the most popular deep learning branches among researchers; in place of Boltzmann, auto-encoding, and sparse encoding method [7]. It is considered an effective classifier and widely used in computer vision. From 2012 to 2014, the AlexNet, Clarifai, and GoogLeNet models were regarded as the best CNN model in ImageNet Large Scale Visual Recognition

Challenge (ILSVRC). Meanwhile, [8] implemented the CNN-AlexNet as a classifier to detect fall events. The CNN-AlexNet was trained and tested using simulated daily activities and fall dataset-Le2i [9] and the result shows a high accuracy rate (99%) of classification. A fall event tracking performance test was performed on CNN model based on KERAS framework [10]. The results showed the sensitivity and specificity rates ranged between 93%-100% and 94%-99%, respectively. According to the study, the use of available CNN models such as AlexNet and KERAS are seen to be able to provide good fall detection performances. Whereas, [11] and [12], proposed a conventional method to detect the human pose and falling event by using a set of polygonal shape features and finite-state machine-fall detection. In other application, a pre-trained ResNEt34 model was implemented on a robust face recognition system in a multi-view vision environment and able to give a good accuracy performance [13].

In this paper, an action recognition system based on CNN-AlexNet deep learning architecture is proposed. The aim is to evaluate the model to classify normal and abnormal actions based on human posture images. The rest of this paper is organized as follows: Section 2 illustrates the proposed methodology in detail. Results and discussion are presented in Section 3. Finally, Section 4 presents the conclusion of this paper.

2. METHODOLOGY

To date, there is no standard definition used by researchers for the term 'events' that focuses on human activity supervision in computer vision. As stated in [14], 'events' is defined as a set of various 'activities' or 'actions' sequences. Whereas, 'activity' is referred to as the repetition of movement [15]. Meanwhile, [16] defined a sequence of 'action' in a certain period as a specific pattern of movement or behavior that can be translated through the human body [17]. As in [18], 'action' is defined as the sub-event of an event. Thus, an event occurring when two different circumstances (activities) resulting from the series of actions in a given place and period. Whereas, the occurrence of an abnormal event is a set of difference activities and actions comparing with set of activities and actions in normal event as illustrated in Figure 1.

This study focuses on detecting a single human action via a single camera view. Training and testing datasets for classifiers were acquired from two different databases: CASIA Gait database [19] and Le2i [9]. The diversities of human posture images were categorized into two groups of action (denoted as normal action (NA) and abnormal action (AA)). The NA image set consists of human actions of normal activities, such as walking and standing as shown in Figure 2(a). On the other hand, the AA image set consists of the action of anomaly fall activities, such as bending, squatting, kneeling, sitting, crawling, and lying down as shown in Figure 2(b).

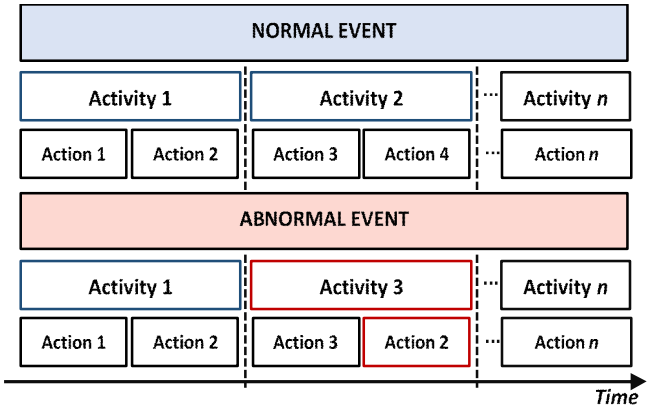

Figure 1: Comparison of normal and abnormal 'events' in term of 'activity' and 'action'

Figure 2: Image samples of normal and abnormal actions from CASIA and Le2i database

Each subject performed the activity independently and it is recorded from a variety of viewpoint angles from multiple stationary cameras. All scenes were taken place inside the building. In this study, the distance between the subject and camera was not determined. However, the entire physical body of the subject was recorded in the field of camera vision.

2.1 Pre-processing

In computer vision, each visual recorded in the form of an image sequence contains information that contributes to detecting human action. However, only the object of interest in the image has to be extracted and thus, simplifying the classifier to extract only the significant features. The external factors such as scattered background conditions and the impact of dynamic lighting changes contained in the image may reduce classification performance. Therefore, background subtraction method is used to segment the moving object [11]. Then, the region of interest (ROI) RGB image is scaled to 224×224 pixels as per input requirement of the CNN-AlexNet model. The pre-processing process is shown in Figure 3.

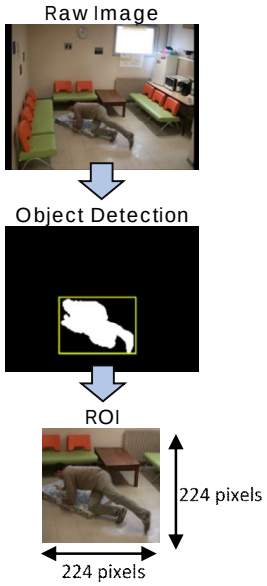

Figure 3: Extracting object of interest from the image

2.2 Action Recognition

In this study, a pre-trained CNN-AlexNet model [20] is used to classify the two-state of action namely, NA and AA. The best CNN model of ILSVRC 2012 had been fully trained with 1.2 million images from the ImageNet database [21] to discover a variety of effective features to classify over 1000 object classes. Therefore, the re-training time for this pre-trained model should be shorter. The basic architecture of CNN for image classification is shown in Figure 4.

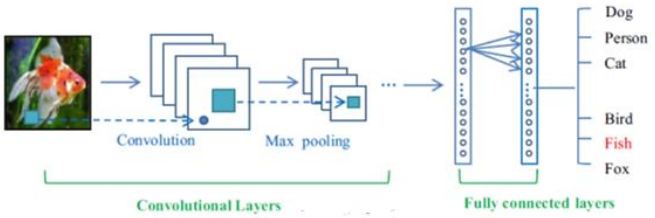

Figure 4: The pipeline of the general CNN architecture [20]

In general, a CNN architecture is divided into two main neural layers, namely the convolution and fully connected layers. In the convolution layer, CNN uses multiple kernels to extract features from the entire image and generates multiple feature mapping. Subsequently, the number of feature mapping dimensions and network parameters is reduced. The fully connected layers perform like a traditional neural network and it contains about 90% of the parameters in a CNN. The layers convert 2D feature mapping to 1-dimensional feature vectors. A large-scale input RGB image should be considered during the training and testing process. The pre-trained CNN-AlexNet model is available at [22].

2.3 Classification Evaluation

Specific performance characteristics are considered such as accuracy (*Acc*), precision (*Pr*), specificity (*Sp*), and F-Score (*Fsc*) rates as derived in (1)-(4):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

$$Pr = \frac{TP}{TP + FP} \times 100\% \quad (2)$$

$$Sp = \frac{TN}{TN + FP} \times 100\% \quad (3)$$

where;
True positives (*TP*) is number of normal actions correctly detected;
False positives (*FP*) is number of normal actions detected as abnormal;
True negatives (*TN*) is number of abnormal actions correctly detected;
False negatives (*FN*) is number of abnormal actions detected as normal.

$$Fsc = 2\frac{Pr \times Rc}{Pr \times Rc} \quad (4)$$

where;

$$Recall\ (Rc) = \frac{TP}{TP + FN} \quad (5)$$

In addition, the duration of training (t_T) and computational (t_C) for classification are also being considered to determine the best classifier.

3. RESULT AND DISCUSION

This section presents the results of bi-action detection using the CNN-AlexNet model. The training and testing dataset of NA and AA which consists of 10,000 images per set were extracted from CASIA and Le2i. A 10-fold cross-validation method was used to test the effectiveness of the machine learning models. The action detection algorithm was built and tested using Matlab® 2017a software on a Windows 10 Home operating system with 2.4 GHz Intel® i7 processor and 16GB RAM.

In general, the classifier showed a good performance with *Acc* and *Fsc* rates were almost 100% as shown in Table 1. Besides, the *Se* and *Sp* rates were high and balanced. However, the training time (t_T) of this CNN model was very high (approx. 8 hours) with the specified computer system. Nevertheless, the computer system required about 0.6s to detect the action in an image. It is anticipated that t_T and t_C can be significantly reduced simply by using a high-performance computer.

Table 1: Performance of CNN-AlexNet model for bi-action detection

Acc	*Fsc*	*Se*	*Sp*	t_T	t_C
99.97%	99.97%	99.94%	100%	8 hours 22 minutes	662.65 ms

4. CONCLUSION

The CNN-AlexNet model has been trained and tested to classify the normal and abnormal actions based on daily human activity and fall images (CASIA and Le2i datasets). The model performed well in detecting the normal and abnormal actions with both *Acc* and *Fsc* rates 99.97%. Unfortunately, with low specifications of a computer hardware system limiting the performance of the classifier with high training time (t_T) and computational time (t_C) (8 hours 22 minutes and 62.65 ms, respectively). However, the CNN-AlexNet model can be considered as a high-performance classifier to recognize actions in overall and adaptable for a smart camera-based surveillance system particularly for monitoring elderly activities in daily living by considering the availability of high computer hardware resources.

ACKNOWLEDGEMENT

The authors acknowledge the financial support from Dip-2018-020.

REFERENCES

1. D. Goldman et al. **Substantial Health and Economic Returns from Delayed Aging May Warrant a New Focus for Medical Research**, *Health Aff.*, Vol. 32, pp. 1698–1705, 2014. https://doi.org/10.1377/hlthaff.2013.0052
2. J. Byles, C. Curryer, N. Edwards, N. Weaver, C. D'Este, and J. Hall. **The Health of Older People in Selected Countries of the Western Pacific Region**, 2014.
3. LPPKN. **5th Malaysian Population and Family Study**, 2014.
4. M. P. Tan, S. B. Kamaruzzaman, M. I. Zakaria, A. V. Chin, and P. J. H. Poi. **Ten-year Mortality in Older Patients Attending the Emergency Department after a Fall**, *Geriatr. Gerontol. Int.*, Vol. 16, pp. 111–117, Jan. 2015. https://doi.org/10.1111/ggi.12446
5. S. Ranasinghe, F. Al Machot, and H. C. Mayr. **A Review on Applications of Activity Recognition Systems with regard to Performance and Evaluation**, *Int. J. Distrib. Sens. Networks*, Vol. 12, Aug. 2016. https://doi.org/10.1177/1550147716665520
6. L. Maddisetti, R. K. Senapati and J.V.R. Ravindra. **Training Neural Network as Approximate 4:2 Compressor applying Machine Learning Algorithms for Accuracy Comparison**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 8, 2019. https://doi.org/10.30534/ijatcse/2019/17822019
7. Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew. **Deep Learning for Visual Understanding : A Review**, *Neurocomputing*,Vol. 187, pp. 27–48, 2016. https://doi.org/10.1016/j.neucom.2015.09.116
8. L. Anishchenko. **Machine Learning in Video Surveillance for Fall Detection**, *Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, 2018, pp. 99-102. https://doi.org/10.1109/usbereit.2018.8384560
9. I. Charfi, J. Miteran, J. Dubois, M. Atri, and R. Tourki. **Optimized Spatio-temporal Descriptors for Real-time Fall Detection: Comparison of Support Vector Machine and Adaboost-based Classification**, *J. Electron. Imaging*, Vol. 22, Jul. 2013. https://doi.org/10.1117/1.JEI.22.4.041106
10. A. Núñez-marcos, G. Azkune, and I. Arganda-carreras. **Vision-based Fall Detection with Convolutional Neural Networks**, *Wirel. Commun. Mob. Comput.*, 2017.
11. M. F. A. Hassan, A. Hussain, and M. H. M. Saad. **Polygonal Shape-based Features for Pose Recognition using Kernel-SVM**, *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, Vol. 10, pp. 41–46, 2018. https://doi.org/10.1155/2017/9474806
12. M. F. A. Hassan, M. H. M. Saad, M. F. Ibrahim, and A. Hussain. **A Finite State Machine Fall Detection using Quadrilateral Shape Features**, *Bull. Electr. Eng. Informatics*, Vol. 7, pp. 359–366, 2018. https://doi.org/10.11591/eei.v7i3.1184
13. J. R. B. D. Rosario. **Development of a Face Recognition System Using Deep Convolutional Neural Network in a Multi-view Vision Environment**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 8, 2019. https://doi.org/10.30534/ijatcse/2019/06832019
14. S. R. Ke, H. Thuc, Y. J. Lee, J. N. Hwang, J. H. Yoo, and K. H. Choi. **A Review on Video-Based Human Activity Recognition**, *Computers*, Vol. 2, pp. 88–131, Jun. 2013. https://doi.org/10.3390/computers2020088
15. T. Ko. **A Survey on Behavior Analysis in Video Surveillance for Homeland Security Applications**, *37th IEEE Applied Imagery Pattern Recognition Workshop*, 2008, pp. 1-8. https://doi.org/10.1109/AIPR.2008.4906450
16. L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra. **Event Detection and Recognition for Semantic Annotation of Video**, *Multimed. Tools Appl.*, Vol. 51, pp. 279–302, 2011. https://doi.org/10.1007/s11042-010-0643-7
17. A. Purohit and S. S. Chauhan. **A Survey on Human Action Recognition**, *IOSR J. Comput. Eng.*, Vol. 19, pp. 43–50, 2017. https://doi.org/10.9790/0661-1904064350
18. M. R. Malgireddy, I. Nwogu, and V. Govindaraju. **A Generative Framework to Investigate the Underlying Patterns in Human Activities**, *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1472–1479. https://doi.org/10.1109/ICCVW.2011.6130424
19. S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan. **Robust View Transformation Model for Gait Recognition**, *18th IEEE International Conference on Image Processing*, 2011, pp. 2073–2076. https://doi.org/ 10.1109/ICIP.2011.6115889

20. A. Krizhevsky, I. Sutskever, and G. E. Hinton. **ImageNet Classification with Deep Convolutional Neural Networks**, *Proceedings - NIPS*, 2012, pp. 1–9. https://doi.org/ 10.1145/3065386
21. Stanford Vision Lab. **ImageNet**, 2016. [Online]. Available: http://www.image-net.org/. [Accessed: 04 August 2019].
22. MathWorks Deep Learning Toolbox Tea. **Deep Learning Toolbox Model for AlexNet Network**, 2018. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange [Accessed: 04 August 2019].