ASIAN JOURNAL OF INFORMATION TECHNOLOGY

MEDWELL PUBLICATIONS

Real-time Burglar Recognition Based on Human Skeletal Data Using Openpose and Long Short Term Memory Network

[1]Shadiya Mohammed Raly and [2]Priyantha Kumarawadu
[1]*Asia Pacific Institute of Information Technology, Sri Lanka*
[2]*ESoft Metro Campus, Sri Lanka*

Key words: Long-short team memory, open pose, real time screaming protocol, 2D skeletal data, burglar recognition

Corresponding Author:
Priyantha Kumarawadu
Asia Pacific Institute of Information Technology, Sri Lanka

Page No.: 1-5
Volume: 21, Issue 1, 2022
ISSN: 1682-3915
Asian Journal of Information Technology

Abstract: The recognition of a burglar caught in CCTV surveillance in real time remains a challenging problem in the domain of action recognition. Existing security systems prioritizes the data acquired from sound sensors, motion sensors, glass breaker sensors over visual sensors to understand the context behind a sequence of action. The proposed system uses the Real-Time Streaming Protocol (RTSP) address of the surveillance camera to acquire the live surveillance images and then uses Open Pose which is a real-time person key point detection library to extract 2D skeletal data which are then fed into a Long-Short Term Memory (LSTM) model, Recurrent Neural Netowrk (RNN) model and Gated Recurrent Unit (GRU) model for classification. The experimental results showed that the performance of LSTM based classifier out performed against RNN and GRU based classifiers under various burglar actions and it was were promising with a training and a validation accuracies of 92.3% and 86.5%, respectively.

INTRODUCTION

Surveillance cameras are used in public places and organizations for indoor and outdoor surveillance in order to establish improved visual security. With the rapid development of emerging technologies, the human operator for monitoring the live surveillance is substituted with machine learning and deep learning. The problem of detecting burglar is identified as a detection of human action which is a standard computer vision problem. Human Action Recognition (HAR) is a popular computer vision problem with a long history and still continues to advance with the improved processing power available. A human activity is determined primarily based on the kinetic state of the subject performing the action. HAR contains a lot of variable factors such as: environment, background lighting, partial occlusion, viewpoint, weather conditions, background noise. According to a survey conducted on different approaches of HAR they divided activity recognition into: (i) Action-based, (ii) Motion-based and (iii) Interaction based activities[1-2]. The pose based approach in action recognition is a popular approach utilized to recognize action in different contexts considering how this is robust against body scales, motion speed, camera viewpoints and partial occlusion. The goal of a pose estimation model is to locate the key points and the associations with the key points[3].

A study utilized locality to determine anomaly from spatiotemporal features extracted utilizing appearance-based approach which takes into account not just the

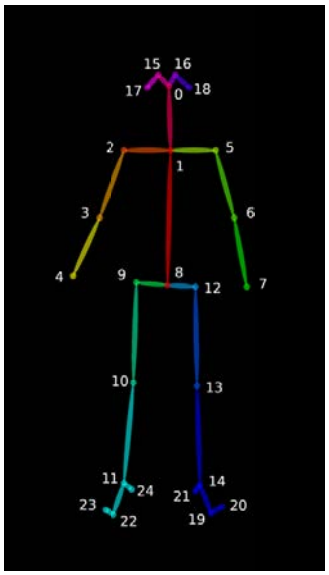

Fig. 1: Pose output format (BODY_25)[6]

subject of concern but also the surrounding. They used a tube extraction module to extract the action tube which is then encoded by a pre-trained 3D network where a regression network followed to classify the behavior[4]. The research group that introduced the University of Central Florida (UCF) crime dataset classified 13 different anomalous categories by introducing a novel multiple instance learning ranking loss for training the model. The research study separated the videos into positive and negative bags based on the action. The anomalous event recognition was assumed as a regression problem and solved by this research[5].

OpenPose which is a real-time pose estimation library which was developed by Carnage Mellon University for realizing posture estimation of human body movement multi-person using 25 joint points of human body as how in Fig. 1. Starting from Point 0 which represents the nose, Point 1 for the neck, 2-7 for shoulders, elbows and wrist of left and right hands, Point 8 for center of hip, Points 9-14 for hips, knees an dancles of left and right legs, Points 15-18 left and right eyes and ears and Points 19-24 for soles, toes and heels of left and right foots have been used for posture estimation. OpenPose posture estimation has been used to action detection by converting data extracted through OpenPose to RGB images in which the motion parameters were encoded in three channels and then by processing them with a neural network for classification[6]. Another study which used OpenPose to extract temporal features classified simple actions such as clapping, jumping, punching by calculating the angle of body joints to determine the motion features for a time duration[7]. Similarly, multiple studies where single action recognition on multiple contexts like yoga, exercises, sports, fall detection have used OpenPose for temporal feature extraction but none of these studies attempted to research the anomality detection in human actions[8-10].

MATERIALS AND METHODS

Despite the change in context of the action to be recognized compare to existing studies, the assumption that spatiotemporal features extracted from the pose-based approach can differentiate burglar from normal behavior was made in this research. The super-imposed sequences of actions exhibited during a normal behavior and a burglar led to the requirement of using an approach which can understand data of 3 dimensions. The assumption of using sliding window algorithm in Long Short Term Memory (LSTM) was made considering the state persistence during each time step of the window by the algorithm. In addition, a Recurrent Neural Network (RNN) model and Gated Recurrent Unit (GRU) model have been trained and tested to benchmark the proposed model.

The proposed LSTM model, RNN model and GRU models were trained in Google Colaboratory with a custom dataset which was created by mixing burglar videos from the existing UCF crime dataset and videos created for this study by us. The spatiotemporal features extracted from OpenPose in real-time from surveillance cameras using Real Time Streaming Protocol (RTSP) [11] are preprocessed before sending them to classification models for classification and if models recognize the sequence as behavior a notification is sent to the application. The overall pipeline of the approach taken is displayed in Fig. 1 and includes streaming live videos using RTSP protocol, Pose-based feature extraction, feature pre-processing and action classification using classification models based on three classifier. In the domain of surveillance security IP cameras and CCTV cameras are the primary options. In this approach the vision field of the camera was tested when camera was placed facing the front door of the house which was selected for the experiment. The angle of the camera was tested in both from an angle of elevation and depression.

Skeletal feature extraction: OpenPose as the feature extractor despite requiring higher computational power comparatively to other pose feature extractors like Alphas Pose or lightweight PoseNet, OpenPose provides support for RTSP and Real Time Messaging Protocol (RTMP) live streaming and has better skeletal feature extraction accuracy which is critical for this solution[12]. In addition, partial skeleton detection during partial occlusion is also supported by OpenPose. OpenPose 25 _BODY model which returns 25 body key points excluding the joints of the face and hand was fed through a video input and the output was written as JSON files with the keypoints in the body part location.

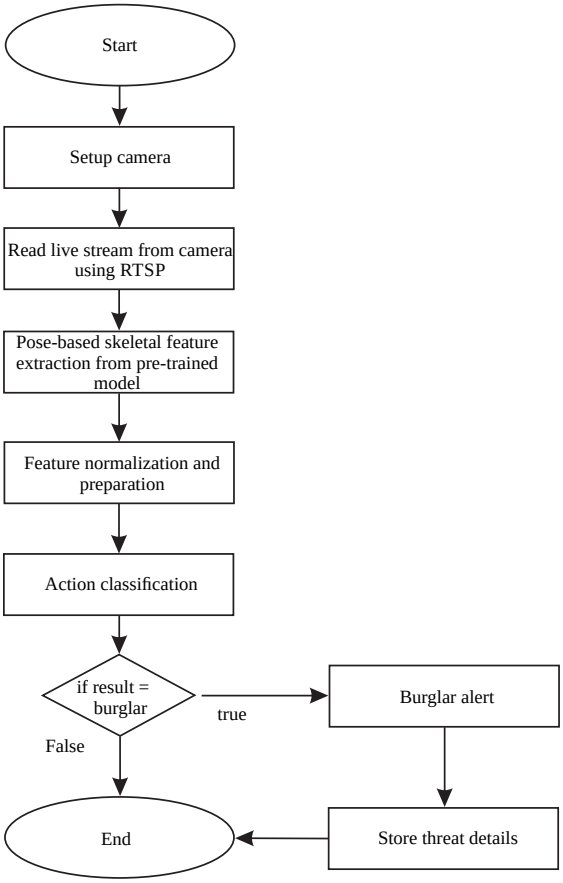

Fig. 2: Overall pipeline of the burglar detection architecture

The JSON file returns the x and y coordinates of each joint and body part and the detection confidence ranged from 0 and 1 depending on the heat map which evaluated the presence of the joint. Some key joints such as left ear, right ear and eyes were not considered in this study since they were not significant for recognition of burglar actions[6]. For each joint, 3 features were needed to be trained and this caused the size of the model to be increased and consequently resuled in increase in the amount of data required to train the model. OpenPose output was preprocessed before feeding them classification. In the preprocessing pipeline, animals of joint positions due to occlusion and inaccuracies due to OpenPose joint assignments were initially removed by using a confidence score as a threshold. Further normalization of data was done by calculating midpoint between hip-joints.

A total of 240 clips labeled as burglar and normal were recorded at 30 fps each with a length of 25 sec and were used in this research. A total of 228,640 frames was acquired from these clips. The only other dataset which matched with our requirement was the UCF Crime dataset. But due to the vast majority of clips being obscured or of varied quality with a mixed of both residential and commercial burglars, we developed a dataset using six subjects. In this study, five common burglar actions were considered: braking a door with a hammer, picklock, banging on the door with the shoulder, kicking a door with a foot and trying to open the door with a lever. The total number of data set was randomly divided into 80% for training and 20% for testing. During the experiment, the model was trained and tested against a varied combination of features in order to acquire the best results. The features of the dataset were evaluated and facial features which included eyes, nose and ears and feet joints such as toes and balm were obscured in a majority of frames returning 0 values. The confidence for approximately 17,000 frames is zero which conclude the idea of utilizing facial features to train the model. According to previous researches which used the skeletal approach to detect a particular action, the use of raw coordinates of the joints did not profit the model[13]. Therefore, the co-occurrence of the joints instead of wholly comparing the features were adopted[14]. The recurrent neural network was developed as a fully connected network where all inputs of the first layer which represented features and outputs of a given layer as the input to the following layer.

Deep learning models: We have used three models: LSTM, RNN and GRU based models in this research and classification results were compared with varied burgler actions. Classification models accepted 3-dimensional data in which batch size, features and time steps were taken into consideration. The models were trained against varied lengths of video with masking in order to analyze and support real world situations and every action exhibited. The models were experimented with both binary cross entropy and mean squared error loss functions which was argued in a previous research in time series problems which treated the problem as a regression.

LSTM has been a popular choice in tackling long term dependencies in researches that involved sentimental analysis, speech recognition and also in behavior recognition[15]. In this research problem where the sequences can have overlapping actions, persistence of previous frames is required to understand the scenario wholly. An example is when a burglar attempts to picklock the sequences will contain the subject focused on the key hole once successful similar to the normal sequence the subject will open the door using the handle and this sub-action of opening the door will cause an overlap between normal and burglar sequences hence a recurrent type of neural network is determined as the most ideal algorithm for this study.

A sequential LSTM model composed of 6 layers were developed in which the first layer was a masking layer

that was used to remove the padding inserted to make the sequences even. The masked value was set as -1, the input number of features is 45 since the facial and foot key joints were removed. Next a Dense layer followed by an LSTM layer which maintained the 3 dimensional input shape. A dropout layer of rate 0.5 followed by a second LSTM layer was then used. Finally, a dense layer with one neuron and the sigmoid activation function as the output layer was compiled with the Adam optimizer with a learning rate of 0.0001 and with mean squared error loss function. The experiments were conducted for varied hyper parameters, dropout rates, the number of neurons an optimum classification results was obtained.

A basic RNN model which was able to learn sequential connection was the second classification model which was implemented. GRU model was built with two GRU layer followed by another with ReLU activation function, A drop out layer was added in between two GRU layers with a probability of 0.2. A fully connected with a softmax activation function was used for classification. Adam optimizer with a learning rate of 0.001 was used during the training.

RESULTS AND DISCUSSION

One of the primary goals in this study was to reduce false alarm rate considering how the existing systems false alarm rate is approximately 94-99%[16]. The accuracy in the approach had a lot of major fluctuations where a lot of modifications were made in order to reach the optimum result in the model. The dataset was cleaned a second time where all the clips were trimmed and analyzed a second time and the malformed and flickering skeletons were removed. The length of the sequences was limited to 350 time steps, anything above it reduced the accuracy of the model. Upon further evaluation similar to existing researches which has experimented time series forecasting problems using LSTM with mean squared error loss function to achieve positive results, this research too achieved augmented results with mean square error compared to a binary cross entropy[17]. An accuracy classification accuracy of 86.5% with an average training accuracy of 92.3% as dipcted in Fig. 3. We represented details of experimental performance of all three models in Table 1 and the proposed LSTM model showed promising results in burglar recognition. The performance testing on the implemented system Upon analyzing the times properly to read the JSON files and make predictions from the model takes only 0.3-0.5 sec hence the problem lied elsewhere. Upon further analysis since OpenPose not only write JSONS but also images, the delay was observed due to this factor. The delay hit an approximate 4 min due to the fact that output images

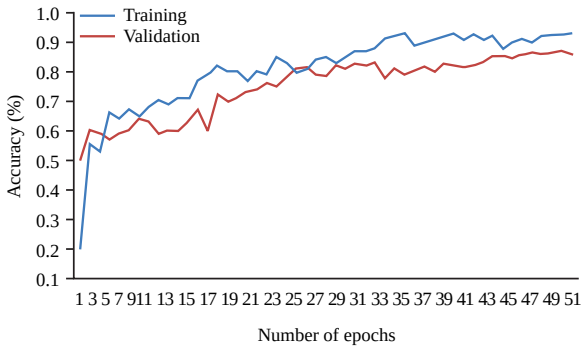

Fig. 3: Model training accuarcy variation gor the LSTM model

Table 1: Classification results

Metrics	LSTM	RNN	GRU
Accuracy	0.865	0.812	0.841
Precision	0.881	0.835	0.862
Recall	0.980	0.831	0.861
F1	0.871	0.826	0.854

of OpenPose drew the skeleton of the subject. The performance was improved to send an alert between 1-2 min by removing this feature.

CONCLUSIONS

A majority of the existing research follow the appearance-based approach for burglar detection and anomalous behavior detection, but after the introduction of pose-based approach it acquired a lot of interest in action recognition and pose correction solutions but none of the researches attempted to utilize the pose-based approach to differentiate burglar actions from a normal sequence. Different pose-based feature extractors were examined during this research and OpenPose was selected based on the accuracy and the fact that OpenPose provides support to IP cameras motivated the decision in selecting OpenPose as the feature extractor. Due to the high end GPU required to run OpenPose the entire solution was implemented in Google Colaboratory. Considering the overall accuracy of the proposed LSTM based burglar recognition model was 86.5% and has outperformed the RNN and GRU models. The current model was trained to determine a single person's action as burglar or normal to extend the model to determine using multi-person skeletons to predict burglars would complicate the current model but with the approach taken since OpenPose does support multi-person pose extraction the extension to evaluate multi persons can be undertaken as a future enhancement. PK and SMR contributed to the conception of the study. SMR performed analysis and the experiment; Both SMR and PK performed the analysis

with constructive discusions and contributed to writing the manuscript. Both authors read and approved the manuscript.

ACKNOWLEDGMENTS

Thanks to everyone who helped with the experiment. We are also very thankful for the editors and anonymous reviewers. This research received no external funding. The authors declare no conflict of interest.

REFERENCE

1. Hussain, Z., M. Sheng and W. E. Zhang, 2019. Different approaches for human activity recognition: A survey. https://arxiv.org/abs/1906.05074
2. Ranasinghe, S., F.A. Machot and H.C. Mayr, 2016. A review on applications of activity recognition systems with regard to performance and evaluation. Int. J. Distributed Sensor Networks, 10.1177/1550147716665520
3. Sun, K., Xiao, D. Liu and J. Wang, 2019. Deep high-resolution representation learning for human pose estimation. Conference on Computer Vision and Pattern Recognition. 2019 IEEE 5686-5696.
4. Landi, F., C.G.M. Snoek and R. Cucchiara, 2019. Anomaly locality in video surveillance. https://arxiv.org/abs/1901.10364
5. Sultani, W., C. Chen and M. Shah, 2019. Real-world anomaly detection in surveillance videos. https://arxiv.org/abs/1801.04264
6. Aubry, S., S. Laraba, J. Tilmanne and T. Dutoit, 2019. Action recognition based on 2D skeletons extracted from RGB videos. Matec. Web. Conf., vol. 277. 10.1051/matecconf/201927702034
7. Noori, F.M., B. Wallace, M.Z. Uddin and J. Torresen, 2019. A robust human activity recognition approach using openpose, motion features and deep recurrent neural network. Lecture Notes Com. Sci., 11482: 299-310.
8. Chen, W., Z. Jiang, H. Guo and X. Ni, 2020. Fall detection based on key points of human-skeleton using openPose. Symmetry, 10.3390/sym12050744
9. Shuai, S., M.S. Kavitha, J. Miyao and T. Kurita, 2019. Action classification based on 2D coordinates obtained by real-time pose estimation. International Workshop on Frontiers of Computer Vision (IW-FCV)At: South Korea. 2019 1-6.
10. Ruether, T., 2022. Streaming protocols: Everything you need to know (Update). https://www.wowza.com/blog/streaming-protocols
11. Cao, Z., G. Hidalgo, T. Simon, S.E. Wei and Y. Sheikh, 2019. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. https://arxiv.org/abs/1812.08008
12. Zhu, W., C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen and X. Xie, 2016. Co-occurrence eature learning for skeleton based action recognition using regularized deep LSTM networks. https://arxiv.org/abs/1603.07772
13. Li, C., Q. Zhong, D. Xie and S. Pu, 2018. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018 Jerome Lang, 786-792.
14. Blackstone, E.A., S. Hakim and B. Meehan, 2020. Burglary reduction and improved police performance through private alarm response. Int. Rev. Law Econ., 10.1016/j.irle.2020.105930
15. Ding, Y., Q. Yang, H. Yu, H. Wang, X. Chen and H. Pu, 2019. Research on real-time behavior recognition method based on deep learning. Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering. 2019 WCSE 307-311.
16. Ammar, S.M.R., M.D. Anjum, R. Islam and M.T. Islam, 2019. Using deep learning algorithms to detect violent activities. http://dspace.bracu.ac.bd/xmlui/handle/10361/12270
17. Riaz, R., S.S. Rizvi, A. Mushtaq, S. Shokat and S.J. Kwon, 2019. Burglar detection using deep learning techniques. J. Eng. Appl. Sci., 14: 2672-2686.