Detecting Harmful Activity in Pilgrimage Using Deep Learning

Musa Dima Genemo
Study Program Computing Software Engineering
Gumushane University, Turkey
Email: musa.ju2002@gmail.com

Abstract—CCTV surveillance is the most extensively used intelligent latest innovation. The use of surveillance cameras has risen dramatically because of the con-venience of monitoring from anywhere and the reduction of crime rates in public areas. In this paper, we introduce the idea of bad vibe activity detection from live videos to enhance the security and safety of pilgrims. The proposed bad vibes activity recognition model is intended to be addressed in the most efficient manner possible using cutting-edge technologies such as TensorFlow and Keras. TensorFlow was chosen because the project could be deployed to a mobile environment in the future with the possibility of extension of other areas such as airport security, bus stain, and public areas that may deserve special attention for security checks. We choose MediaPipe Ho-listic for employee bad vibe recognition in the model.

Keywords—Artificial Intelligence, Classification, Real-Time Object Recognition, Computer vision.

Abstrak—Pengawasan CCTV adalah inovasi cerdas terbaru yang paling banyak digunakan. Penggunaan kamera pengawas telah meningkat secara dramatis karena kemudahan pemantauan dari mana saja dan pengurangan tingkat kejahatan di tempat umum. Dalam makalah ini, kami memperkenalkan ide deteksi aktivitas getaran buruk dari video langsung untuk meningkatkan keamanan dan keselamatan jemaah. Model pengenalan aktivitas getaran buruk yang diusulkan dimaksudkan untuk ditangani dengan cara seefisien mungkin menggunakan teknologi mutakhir seperti TensorFlow dan Keras. TensorFlow dipilih karena proyek dapat diterapkan ke lingkungan seluler di masa mendatang dengan kemungkinan perluasan area lain seperti keamanan bandara, noda bus, dan area publik yang mungkin memerlukan perhatian khusus untuk pemeriksaan keamanan. Kami memilih MediaPipe Ho-listic untuk pengenalan getaran buruk karyawan dalam model.

Kata Kunci—Artificial Intelligence, Klasifikasi, Real-Time Object Recognition, Computer vision.

I. INTRODUCTION

The use of surveillance cameras has risen dramatically because of the conven-ience of monitoring from anywhere and the reduction of crime rates in public areas. Hajj is one of the Five Pillars of the Islamic religion where the pilgrimage to the holy city of Mecca in the kingdom of Saudi Arabia, which takes place in the last month of the year (Hijri calendar) and which all Muslims are obligated to make at least once throughout their lifetime if they can afford it [1]. Before COVID_19 emerged, 2.5 million people would travel every year to Saudi Arabia for Hajj. Due to this, the security of pilgrims needs special attention. New cut-ting-edge technology is required to ensure the safety of the people and the city where the hajj imitation takes place, as well as the detection of forbidden activi-ties and the carrying of prohibited things such as guns, flames, sharp metals, and the like. Human activity recognition (HAR) is the ability to use sensors to ana-lyze human body indicators or motion and identify human actions or events [2]. HAR is regarded as a significant component in various scientific research set-tings, such as health [3], Human-robot interaction [4], and security [5].

Such technologies are in high demand during the hajj festival to safeguard pil-grims' safety. Many people have become victims of the Hajj scam in recent years, losing money, cellphones, and other valuables. Nowadays Terrorist acts pose the greatest danger to public safety [6]. Prohibited things, such as carrying a gun, hurling a bomb, deceiving people, and threatening a suicide bombing, should be checked instantly.

As a result, these challenges demand models that generate a warning or alarm. If accurate forecasts are provided in a timely manner, human lives can be saved by employing this newly introduced model. Interleaved actions, such as throwing a stone at three walls (Ramy Al Jamarat), which is also known as stoning the devil (sheytan) and running between mina and muzdalifah are a pillar of the Hajj pil-grims. The stoning of the devil may cause prediction ambiguity, by throwing stones at people or away from the road and running between mina and muzdali-fah may cause prediction ambiguity with a sudden run. A recurrent Neural Net-work (RNN) is used to overcome activity overlapping difficulties.

Despite this, utilizing smart CCTV surveillance reduces labor expenses while also increasing the security and safety of pilgrims. This study proposes a deep feature extraction mechanism for forbidden motion and activity identification to address the difficulties. We proposed a new model named l4-branched-action net. By using this new model, we extract features from the video frame and bels the activity to activity to their respective class like the need for special attention, or safe move. 64 layers of CNN- deep architecture are used for feature extraction. To optimize the deep features that have been obtained, an ACO feature selection technique is applied. By running convolution layers over pre-trained public data like the CIFRA-100.

II. METHOD

The proposed model will be presented in its entirety in this section. Further-more, this section includes details of the proposed 64-layer Classification algo-rithm. We used the CIFAR-100 dataset to train the proposed model, as well as feature extraction from the action recognition dataset using the proposed CNN architecture, feature selection using Ant Colony Optimization (ACO), and predic-tion using a variety of algorithms. For autonomous feature extraction from video frames and classifications events in the frame, a novel proposed 64-layer CNN architecture is used. The recommended L4-BranchedActionNet's physical archi-tecture is shown in Fig.3 and Fig.4.

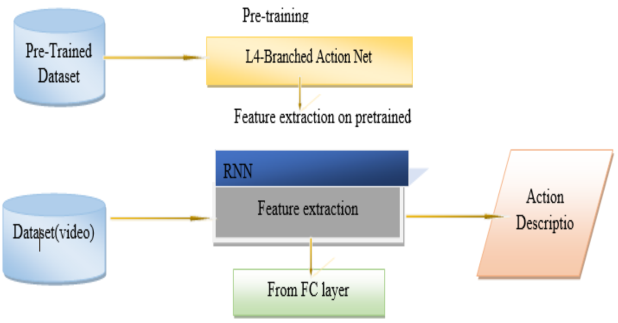

Fig.3. Structure of proposed model

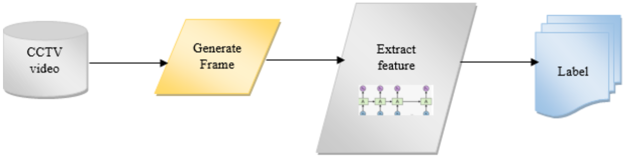

Fig 4. video frame generation

Table 1. Layer configuration of L4-branched action net

Layer #	Layer name	Feature maps	Filter depth	Stride
1	Input	227 × 227 × 3		
2	Conv_1	55 × 55 × 96	11 × 11 × 3 × 96	**[4 4]**
3	ReLU_1	55 × 55 × 96		
4	Batch_Norm_3	55 × 55 × 96		
…..	FC_20	1 × 1 × 100	[1 1]	**Same**
62	Prob	1 × 1 × 100		
63	FC_21	1 × 1 × 100	[1,1]	**same**
64	Video description			

The data was collected using a script generated utilizing OpenCV and MediaPipe Holistic, as shown in Fig.5 frames of data are recorded for each word caught.

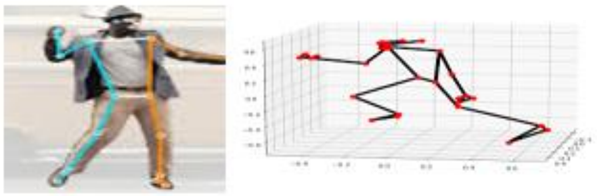

Fig.4. Key using Open Pose using MediaPipe Holistic point extraction [12]

NumPy array is used instead of pictures to hold video frames. We passed three major steps to train the model. The following are the details of the new model's operations steps. The first step the is Conv layer; (1) In the Conv layer the input x i−1 filter is computed using equation 1.

$$Xp^{,j} = fi\left(\sum_p fi,p^{,p} \; * \; xp,i-1+fp^{\hat{}},i\right) \qquad (1)$$

where pj input channels and p^j represent the number of output channels. j represents several layers in the mode, fi filter. Equation 2 is used to calculate the max pool in the pooling layer.

$$Xp,j,u,v = \; maxl = 1..s,m = 1..tXp,j-1,(u+1)(v+m) \qquad (2)$$

where u, v represents the matrix index of frame Xp, j-1and l, m matrix index of the pooling window. It calculates the mean and variance in fragments. The mean is derived, and the features are separated using the standard deviation as follows.

$$\mu = \frac{1}{w\sum_{z}^{w} Xz} \qquad (3)$$

where w is the number of feature maps in a batch. We used both ReLU and Leaky_ ReLU in the proposed model. All numbers less than 0 are transformed to 0 by the standard ReLU, which is stated as [15]:

$$Xu,v = \max\,(0, Xu,v) \qquad (4)$$

For values less than zero, Leaky ReLU has a small slope rather than zero. A leaky ReLU will have v = 0.01u when u is negative. CNN can further be learned in-depth from several works [16-19].

The second step is feature extraction;(2) For feature extraction from a video frame, the appropriate frame is retrieved. The proposed approach is intended to feature extraction from the deep-trained CNN pipeline. We trained the new model on public dataset t such as CIFAR100 [70] which contained images of 1000 classes. The trained network is then used for feature extraction on action recognition datasets and the FC_18 layer is chosen for features extraction. A total of 4096 features is attained per frame from the FC_18 layer. The prepared dataset contains a total of 13250 video frames. This makes the feature set dimension of all datasets 13250 × 4096. Figure 5 illustrates the visualizations of the strongest feature maps at various convolution layers on L4-Branched-ActionNet.

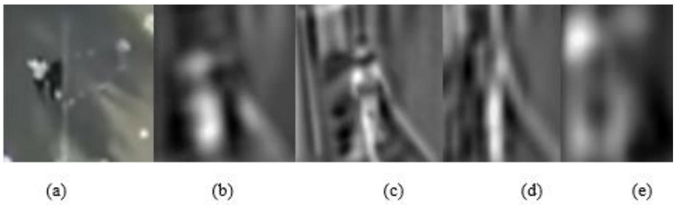

Fig.5. Image visualizations of strongest feature maps at various convolution layers (a) Conv_1, (b) Conv_2, (c) Conv_5, (d) G_Conv_8, (e) Conv_10.

And the third step is (3) after interpreting the received result the extracted features are coded by applying entropy-coded ACO optimization operation [25] using equation (5).

$$e(X1 \dots\dots Xn) = \sum f1 \dots \sum fn(p(f1 \dots fn)LOGp(f1..fn) \qquad (5)$$

Where (x1-xn) represents the feature. We used ACO for feature optimization based on the likelihood at a given point at a certain time. The last step is classification, in which ACO-based chosen features are at the end passed to the predictor for categorization. Several SVM and KNN versions are used to assess model performance. Cub-SVM emerges as the most effective as shown in table 2.

Table 2. Performance of the model

Classifier	Sensitivity	Specificity	Precision	Measure	Percent
LSVM	83.38	72.62	39.94	52.52	77.74
QSVM	89.11	91.53	61.79	76.01	86.14
FGSVM	57.29	51.78	25.02	32.80	54.39
MGSVM	90,52	92.35	62.56	76.58	86.28
CGCVM	68.47	64.45	31.80	41.75	66.33
CSVM	96.33	95.59	76.61	88.08	92.99

For testing, we employed random selection using sklearn's train test function. Following that, Keras' Callback functions were used to improve the training's efficiency. The accuracy of the test data is evaluated. We also used the public dataset ON WEIZMANN to compare our results to the current state of the art. The outcome is shown in table 3.

Table 3. Performance evaluation on weizmann dataset

Method reference	Year	Accuracy
DWT+KNN [21]	2020	0.93
CNN+ELM [22]	2020	0.94
Gabor-Ridgelet Transform [23]	2020	0.93
LCF + MSVM [22]	2021	0.95
ANN [24]	2020	0.80
PCANet-XY-YT [25]	2021	0.91
Ours (L4-Branched-ActionNet + EntACS + Cub-SVM)	-	0.93

III. RESULTS AND DISCUSSION

In The major goal of this study is to develop a CNN architecture that can recog-nize harmful actions during the Hajj festival. Then, the Deep L4-BranchedActionNet Deep Network proposed here is used to extract powerful features. The pretraining is carried out using a publicly available dataset, CIFAR-100. For testing, we employed random selection using sklearn's train test func-tion. Following that, Keras' Callback functions were used to improve the train-ing's efficiency to complete this design, many methods such as fine-tuning, add-ing and removing layers, and neurons were used. Finally, the 64-layer architec-ture was proven it is the most efficient in terms of performance. Tensor flow Keras, OpenCV, and the NumPy library were used in all the experiments in this.

Table 4. confusion matrix of csvm classifier

Sudden ran	0.92021	0.00	0.01	0.02
Fighting	0.00	0.91221	0.00	0.01
Throwing	0.01	0.01	0.90021	0.00
Robbing	0.00	0.00	0.01	0.91002
	Sudden ran	Fighting	Throwing	Robbing

IV. CONCLUSION AND RECOMMENDATIONS

Detection of harmful vibes is critical for pilgrims' safety. To detect banned actions during the hajj festival, we utilized a 64-layer CNN network called L4-Branched-ActionNet. The model is evaluated on datasets that are freely availa-ble, such as the CIFAR-100 object detection dataset. The characteristics were retrieved and subsequently reduced using an entropy-coded ACO. To evaluate model performance, several SVM and KNN versions are utilized. With an accu-racy of 0.91221, Cub-SVM emerges as the most effective. This work will be im-plemented on security personnel's mobile phones for convenient monitoring from any location in future work.

REFERENCES

1. R. K. Tripathi, A. S. Jalal, and S. C. Agrawal, "Suspicious human activity recognition: a review," Artificial Intelligence Review, vol. 50, pp. 283-339, 2018
2. A. Tapus, A. Bandera, R. Vazquez-Martin, and L. V. Calderita, "Perceiving the person and their interactions with the others for social robotics–a review," Pattern Recognition Letters, vol. 118, pp. 3-13, 2019.
3. A. Ilidrissi and J. K. Tan, "A deep unified framework for suspicious action recognition," Artificial Life and Robotics, vol. 24, pp. 219-224, 2019.
4. Konstantinidis, D., Dimitropoulos, K., & Daras, P. (2018). Sıgn Language Recognıtıon Based On Hand And Body Skeletal Data. 2018- 3DTV-Conference: The True Vision - apture, Transmission and Display of 3D Video (3DTVCON).
5. S. J. Elias, S. M. Hatim, N. A. Hassan, L. M. A. Latif, R. B. Ahmad, M. Y. Darus, and A. Z. Shahuddin, "Face Recognition Attendance System Using Local Binary Pattern (LBP)," Bulletin of Electrical Engineering and Informatics, vol. 8, 2019.
6. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks,"

Advances in neural information processing systems, vol. 25, pp. 1097-1105, 2012

7. Genemo, M. D. (2022). Suspicious activity recognition for monitoring cheating in exams. Proceedings of the Indian National Science Academy, 1-10.
8. C. A. Devine and E. D. Chin, "Integrity in nursing students: A concept analysis," Nurse education today, vol. 60, pp. 133-138, 2018.
9. H. M. Abdulghani, S. Haque, Y. A. Almusalam, S. L. Alanezi, Y. A. Alsulaiman, M. Irshad, et al., "Self-reported cheating among medical students: An alarming finding in a cross-sectional study from Saudi Arabia," PloS one, vol. 13, p. e0194963, 2018.
10. M. A. Lewis and C. Neighbors, "An examination of college student activities and attentiveness during a web-delivered personalized normative feedback intervention," Psychology of Addictive Behaviors, vol. 29, p. 162, 2015.
11. RWTH-PHOENIX-2014-T veri seti, https://wwwi6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX- 2014-T/
12. S. Balocco, M. González, R. Ñanculef, P. Radeva, and G. Thomas, "Calcified plaque detection in IVUS sequences: Preliminary results using convolutional nets," in International Workshop on Artificial Intelligence and Pattern Recognition, 2018, pp. 34-42
13. Y. Liu, X. Wang, L. Wang, and D. Liu, "A modified leaky ReLU scheme (MLRS) for topology optimization with multiple materials," Applied Mathematics and Computation, vol. 352, pp. 188-204, 2019
14. J. Bouvrie, "Notes on convolutional neural networks," Neural Nets, MIT CBCL Tech Report, pp. 47-60, 2006.
15. Y. Li, Z. Hao, and H. Lei, "Survey of convolutional neural network," Journal of Computer Applications, vol. 36, pp. 2508-2515, 2016.
16. A. Divakaran, Q. Yu, A. Tamrakar, H. S. Sawhney, J. Zhu, O. Javed, et al., "*Real-time object detection, tracking and occlusion reasoning,"* ed: Google Patents, 2018.
17. A. Booranawong, N. Jindapetch, and H. Saito, "*A system for detection and tracking of human movements using RSSI signals*," IEEE Sensors Journal, vol. 18, pp. 2531-2544, 2018.
18. A. B. Mabrouk and E. Zagrouba, "*Abnormal behavior recognition for intelligent video surveillance systems: A review*," Expert Systems with Applications, vol. 91, pp. 480-491, 2018.
19. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images (Technical Report)," University of Toronto, 2009
20. D. K. Vishwakarma, "*A two-fold transformation model for human action recognition using decisive pose*," Cognitive Systems Research, vol. 61, pp. 1-13, 2020.
21. M. A. Khan, Y.-D. Zhang, S. A. Khan, M. Attique, A. Rehman, and S. Seo, "*A resource conscious human action recognition framework using 26-layered deep convolutional neural network*," Multimedia Tools and Applications, pp. 1-23, 2020.