2021 International Conference on
Artificial Intelligence (ICAI)
Islamabad, Pakistan, April 05-07, 2021

Vision-Based Real-time Human Malicious Behavior Detection

Hajra Binte Naeem
Swarm Robotics Lab, NCRA
UET Taxila, Pakistan
hajra.naeem@uettaxila.edu.pk

Muhammad Haroon Yousaf
Swarm Robotics Lab, NCRA
UET Taxila, Pakistan
haroon.yousaf@uettaxila.edu.pk

Farhan Hassan Khan
KDRC, EME-NUST
Pakistan
mrfarhankhan@gmail.com

Amanullah Yasin
Swarm Robotics Lab, NCRA
UET Taxila, Pakistan
amanyasin@gmail.com

2021 International Conference on Artificial Intelligence (ICAI) | 978-1-6654-3293-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAI52203.2021.9445205

***Abstract*—Human detection and behavior analysis from surveillance videos is an active area of research in computer vision. Authorities and security administrators need a system that can detect human malicious behavior to take immediate necessary actions. In this paper, we propose an approach to detect the anomalous/malicious behavior of humans in the surveillance videos. The proposed approach models the human behavior using human joint motion information from skeleton sequence. We have divided the proposed approach into four sub-modules i.e. human detection and skeleton estimation, human ID assignment, feature extraction and classification. The proposed approach is evaluated on publicly available CASIA dataset in offline mode and accuracy of 90.81% has been achieved. The experimental results indicates that it can be exploited in real-time applications with low computational cost of 18 frames per second.**

***Index Terms*—Computer Vision, Human Behavior, Malicious Activity, CASIA**

I. Introduction

Human behavior analysis from videos is one of the extensively explored areas of research in computer vision. The key objective is to detect, recognize and distinguish human behavior with respect to respective environment and application. Particularly malicious event detection, irregular behavior, unusual activity, anomaly and abnormal behavior is important to detect for indication of any significant incidence. So that authorities and security personnel can take precautionary measures immediately for safety of individuals and property. The applications of malicious behavior detection can be found in environment monitoring, perimeter security, ATM surveillance and other security systems. There is need of intelligent and automated approaches that can strengthen the surveillance systems to minimize the human labor for indication of criminal and accidental incidents. Therefore, the surveillance systems need to be intelligent,automated and real-time responsive.

The basic-level surveillance systems have self-monitoring ability to detect and track target human. The extended ability is to recognize malicious behavior and generate alarms upon the occurance of any such incident. To detect human, [1] proposed an algorithm based on strong correlation between time domain difference image boundary and gray image boundary. [2] combined moving human boundaries extracted by these two images to obtain more precise target boundaries. These methods are based on background subtraction and sensitive to change in motion scene [3]. To overcome the problem, [4] proposed dynamic background update. Moreover, the problem to reduce computational cost of optical flow field is explored to meet real-time requirements [5]. [6] included prior knowledge to detect the targeted human.

Human behavior can be defined as in terms of pose as combination of multiple joints and body movements. Basic behavior recognition algorithm uses frame-to-frame matching with feature sequence [7]. Template matching is simple to compute, but sensitive to execution time. Gruber et. al. [8] proposed context independent dictionary analysis to detect and recognize long-term human behavior and interactions. Traditional approaches extract hand-crafted local features and descriptors such as SIFT [9], HOG/HOF [10] and Extended SURF [11]. These approaches use human body part appearance and movement information.

Current methods [12], [13] use real-time skeleton estimation frameworks that detect body joints or keypoints. Openpose is one of the framework that detects body, face, hands and foot joints. Its major benefit is that the estimation speed is not sensitive to number of people, as compared to MASK-RCNN [14], Fastpose [12] and Alphapose [15], [16]. Fastpose [12] is a top-down approach that first detects each human and than estimates their body joints. But when people are in close proximity human detector may fail and there is no recourse to recovery. And complexity of detection and pose estimation increases with increase in number of humans. Alternatively, Openpose is a bottom-up approach that take clues of body joints from feature maps. These body joints are connected together to estimate complete skeleton of human. Such approach decouples runtime complexity with increase in number of humans.Such approach can solve behavior recognition problem in real-time and is invariant to camera view-point. Skeleton-based behavior recognition is a time-series problem. The skeletal data comprises 2D coordinates of the key joints in the human body over time. So we can represent human actions as the movement of skeleton sequences. Our goal is to exploit the potential of low-cost skeleton estimation algorithm and then explore them to recognize human malicious behavior.

In this paper, we propose an approach an intelligent approach to detect human malicious behavior in CCTV videos. The human joint motion information is incorporated in the proposed approach. The propsoed approach comprised of four sub-modules namely human detection and skeleton estimation,

978-1-6654-3293-1/21/$31.00 ©2021 IEEE

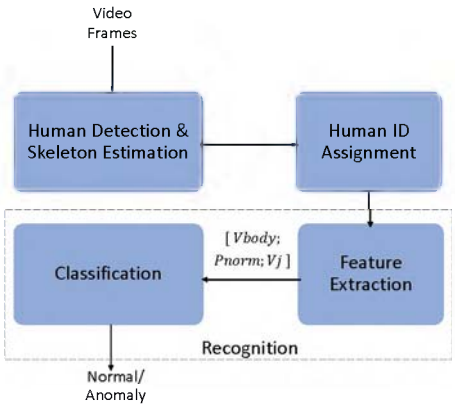

Fig. 1. Proposed method sub-modules

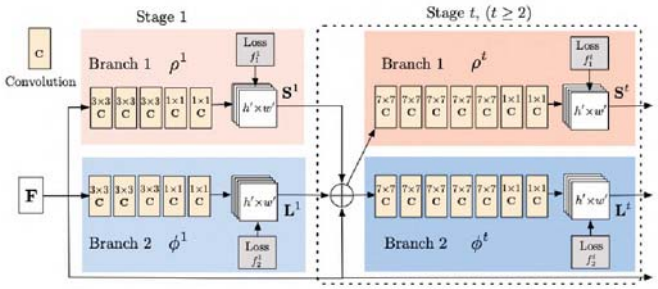

Fig. 2. OpenPose two branch multi-stage CNN architecture [13]

human ID assignment, feature extraction and classification. First two modules detect humans and assign ID and last two recognize activity to distinguish malicious behavior in the respective frames. These submodules are discussed in next section. Each module works independently forming a pipeline as shown in figure I. The system takes RGB video frames as an input and detects anomalous frames at output. The approach has a key advantage in terms of utilization and implementation in real-world scenario for detection of malicious or criminal behavior in CCTV footages.

II. Proposed Approach

A. Human Detection and Skeleton Estimation

The first step in human malicious behavior detection is to detect humans and estimate their skeletons from video frames. Estimated skeleton is approximate 2D pose of human locating each body joint position and their connection. Body joints are the keypoints of a 2D human pose.

In this study, we have used efficient bottom-up OpenPose [13] framework for human detection and skeleton estimation. First feature map F is detected from video frame of size $(w \times h)$ using the first 10 layers of VGG-19. Then two branch multi-stage feedforward convolutional network predicts body joints position and connection, as shown in figure II-A. First branch (ρ) produces a set of 2D confidence maps S of body joint locations. Where, each map is a 2D representation of

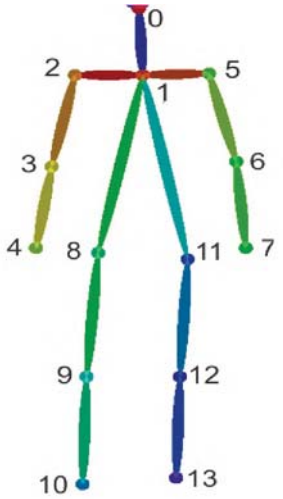

Fig. 3. Human skeleton joint labels

particular body part pixel location approximation. The set $S = (S_1, S_2, S3, ..., S_J)$ includes J confidence maps, one for each body joint represented as, $S_j \epsilon \mathbb{R}^{w \times h}, j \epsilon \{1...J\}$. Second branch $(\emptyset)$ produces a set of 2D vector fields L of part affinities. Part affinities is association between body joints to represent position and orientation of body joint pair. The set $L = (L_1, L_2, L3, ..., L_C)$ includes C vector fields, one for each body joint pair represented as, $L_c \epsilon \mathbb{R}^{w \times h \times 2}, c \epsilon \{1...C\}$.

Loss functions f determine whether the network is convergent at each stage from equation 1.

$$f = \sum_{t=1}^{M} f_1^t + f_2^t \tag{1}$$

where f_1^t and f_2^t are each branch loss function and M is total number of stages. Finally, bipartite [17] graph matching is used to build 2D pose from joint keypoints and edges. The Hungarian algorithm is used to optimally match adjacent nodes, instead of adopting global optimization.

B. Human ID Assignment

Frame spatial information is used to detect 2D pose of human. Now, video temporal information is exploited to record human motion across frames. Technically it is change in body joints positions with respect to previous frame. We need individual human ID, as there could be multiple humans in each frame. Multiple human body joints are tracked to assign human IDs using Simple Online and Real-time Tracking (SORT) [18], which exploits recursive Kalman filtering and frame-by-frame data association. Individual human skeleton sequence H in N frames, denoted by $H = \{P_j^n\}$ where $n = 1, 2, 3, ..., N$ and $j = 1, 2, 3, ..., J$. Where P_j^n is x an y coordinates of j pose in t frame.

C. Human Action Recognition

Next step is to recognize action using pose features. As shown in figure II-B, total 13 joints i.e 1 neck, 3 left arm, 3 right arm, 1 left thigh, 2 left leg, 1 right thigh, 2 right leg

are used for feature extraction. Thus J is set to 13. Three type of features are extracted from human skeleton (1) body velocity $Vbody_j^n \epsilon \mathbb{R}^{1\times 2}$ and (2) normalized joint positions $Pnorm_j^n \epsilon \mathbb{R}^{J\times 2}$ and (3) joint velocities $V_j^n \epsilon \mathbb{R}^{J\times 2}$. Neck joint P_0^n is set as an origin and height h of pose is a Euclidean distance from neck to thigh. Body velocity along x and y co-ordinate is displacement of origin with respect to previous frame. Normalized joint positions are calculated by removing offset of each joint from origin and dividing over mean height. Equation 2 represents normalized position of joint j at frame n.

$$Pnorm_j^n = \left(P_j^n - P_0^n\right)/h \tag{2}$$

Joint velocities along and x and y coordinates are calculated as displacement of joint with respect to previous frame. Equation 3 represents velocity of joint j in frame n.

$$V_j^n = ||Pnorm_j^n - Pnorm_j^{n-1}||_2 \tag{3}$$

Features from sequence of frames are extracted and concatenated to form a feature vector. Length of a feature vector from N frames is $(1 + J + J) \times 2 \times (N - 1)$. These features are then used to train a binary classifier. In this study, we have tested five popular classifiers: K-nearest neighbour(K-NN), SVM with linear kernel, SVM with rbf kernel, decision tree, random forest and deep neural net (DNN) of 3x3x3 for more insightful evaluation. For DNN, rectified linear unit (ReLu) is selected as an activation function. Comparison of classifiers result is given in section III-C.

III. Experimental Evaluation

A. Dataset

CASIA action dataset [19] is composed of human action sequences of outdoor video cameras. The videos are captured from three different view angles. Horizontal and top-down view captures action such that human side-view and head-view are only visible, respectively. Angle-view captures video from camera mounted at 30 degree and most of the human body parts are visible; the couple of these views are more-likely similar to CCTV footages. Total of 1446 sequences are present in the dataset that contains eight types of actions of single person (walk, run, bend, jump, crouch, faint, wander and punching a car). Each action is performed by 24 subjects. The dataset also contains the two person interactions of seven types (rob, fight, follow, follow and gather, meet and part, meet and gather, overtake) performed by every 2 subjects. Actions: walk, run, jump and wonder are labelled as normal. Whereas actions: bend, crouch, faint and punching a car are labelled as anomaly. Only angle-view and horizontal view single person actions are used for training and testing. As OpenPose model cannot detect pose from top-down view, as neck is not visible in such cases.

B. Implementation Details

Original OpenPose is implemented in C using Caffe library. However, to work in python we have used Tensorflow library. OpenPose model 'cmu'is used which is pre-trained on COCO database. It detects 18 human joints (head, neck, shoulder, arms, body and legs) excluding hand and feet joints. Figure III shows skeleton estimation of multiple actions. Human pose features are extracted from non-overlapping five frames (every 0.5s). Data is split into 70-30 training-testing split.

TABLE I
Comparison of classifiers result on CASIA database

Classifier	Class	Precision	Recall	F1-score	Accuracy(%)
K-NN	Normal	0.86	0.98	0.92	85.76
	Anomaly	0.85	0.37	0.52	
SVM-linear	Normal	0.90	0.97	0.93	88.55
	Anomaly	0.81	0.58	0.67	
SVM-rbf	Normal	0.91	0.98	**0.94**	90.79
	Anomaly	0.88	0.64	0.74	
Decision Tree	Normal	0.90	0.95	0.93	88
	Anomaly	0.76	0.61	0.68	
Random Forest	Normal	0.92	0.97	**0.94**	**90.81**
	Anomaly	0.85	0.67	**0.75**	
DNN	Normal	0.92	0.92	0.92	85.57
	Anomaly	0.69	0.70	0.69	

C. Results and Discussion

Table I shows the accuracy of proposed approach with different classifiers. Malicious behaviour detection is imbalanced classification problem. As occurance of anomaly as compared to normal behaviour is rare in long video sequences. Therefore, table I reports per-class precision, recall and f1-score of both classes, which are defined in equation 4, 5 and 6.

$$Precision = P = \frac{TP}{TP + FP'} \tag{4}$$

$$Recall = R = \frac{TP}{TP + FN'} \tag{5}$$

$$F1 - score = \frac{2 \times P \times R}{P + R} \tag{6}$$

where True Positives (TP) represent the number of actions correctly recognized, False Positives (FP) represent the number of actions falsely recognized, and False Negatives (FN) represent the number of missed actions. F1-score is harmonic mean of recall and precision and is better metric for imbalanced classification. Lower value of precision and recall for anomaly shows that number of false negatives and false positives are high. SVM with rbf kernel and random forest has shown best performance with maximum accuracy of 91% approximately. Figure III illustrates per-class f1-score and accuracy comparison of classifiers. Figure III shows results on video frames of anomalous action 'fight'. This action includes interaction of human, whereas model is only trained for individual human skeleton. Where first frame is normal when human are only walking and last two frames are anomalous when human are fighting. Anomaly is detected due to change in body pose and joint features of individual human.

Proposed approach reports average run-time complexity of 18 frames per second. It is number of video frames processed in a second. Our approach detects pose in real-time followed

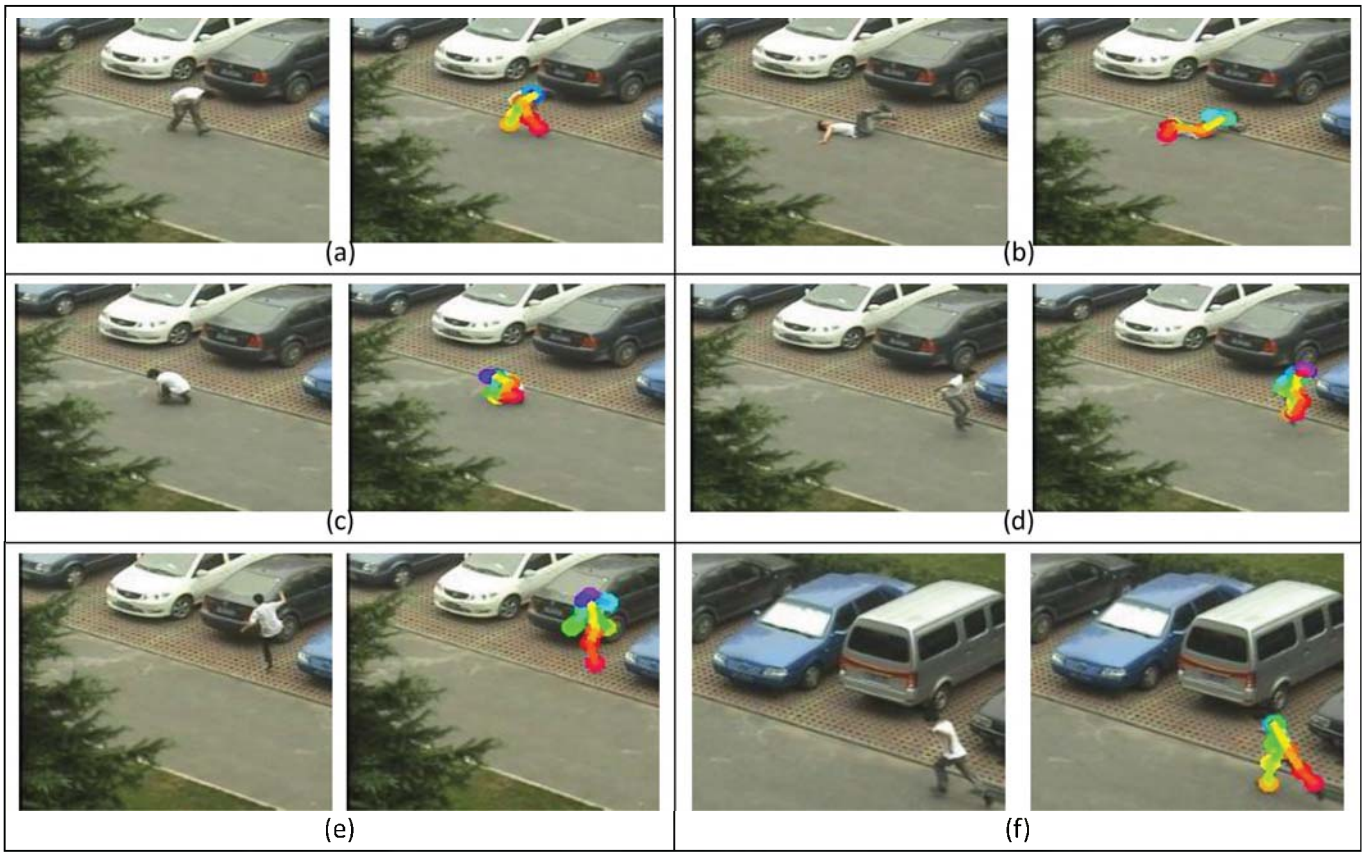

Fig. 4. Skeleton estimation in CASIA database of actions (a)bend (b)faint (c)crouch (d)jump (e) and (e)run

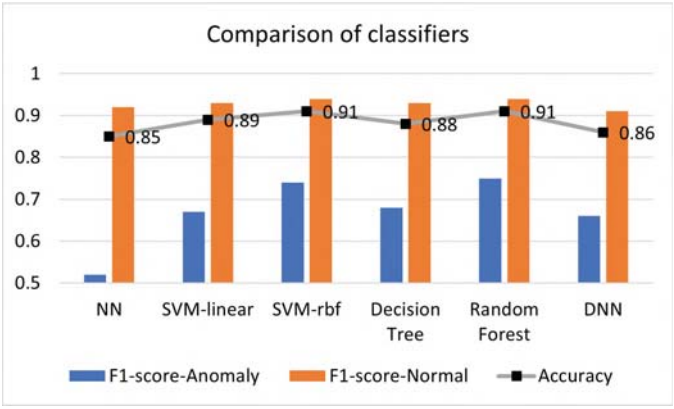

Fig. 5. Comparison of classifiers result

by supervised learning based action recognition. Thus proposed system low computational complexity suggests that it can be implemented in real-time applications.

IV. Conclusion and Future Work

The proposed human behavior detection approach shows significant accuracy and potential for implementation in real-time scenario. Human skeleton extraction using OpenPose and simple joint motion features proved effective for human

Fig. 6. Malicious behavior detection in action 'fight'

behavior detection. The evaluation results on CASIA dataset and the computational cost indicated the implementation in realistic scenario. Keeping in view the simple but effective approach, it can automate the malicious behavior detection process in surveillance systems and may reduce the additional human labor required for this purpose. In future, more complex features and robust representations will be explored for realistic dynamic environments.

ACKNOWLEDGMENT

This research work is supported and funded by Higher Education Commission, Pakistan, for Swarm Robotics Lab under National Centre for Robotics and Automation (NCRA).

REFERENCES

[1] C. Guo, K. Kidono, R. Terashima, and Y. Kojima, "Humanlike behavior generation in urban environment based on learning-based potentials with a low-cost lane graph," *IEEE Transactions on Intelligent Vehicles*, vol. 3, no. 1, pp. 46–60, 2017.

[2] J. Zhang, J. Geng, J. Wan, Y. Zhang, M. Li, J. Wang, and N. N. Xiong, "An automatically learning and discovering human fishing behaviors scheme for cpscn," *IEEE Access*, vol. 6, pp. 19 844–19 858, 2018.

[3] L. E. Wylie, K. P. Hazen, L. A. Hoetger, J. A. Haby, and E. M. Brank, "Four decades of the journal law and human behavior: a content analysis," *Scientometrics*, vol. 115, no. 2, pp. 655–693, 2018.

[4] L. Liu, S. Ma, L. Rui, and J. Wen, "Locality constrained dictionary learning for human behaviour recognition," *Journal of Statistical Computation and Simulation*, vol. 87, no. 13, pp. 2526–2537, 2017.

[5] M. Devanne, S. Berretti, P. Pala, H. Wannous, M. Daoudi, and A. Del Bimbo, "Motion segment decomposition of rgb-d sequences for human behavior understanding," *Pattern Recognition*, vol. 61, pp. 222–233, 2017.

[6] T. P. Almeida, G. S. Chu, M. J. Bell, X. Li, J. L. Salinet, N. Dastagir, J. H. Tuan, P. J. Stafford, G. A. Ng, and F. S. Schlindwein, "The temporal behavior and consistency of bipolar atrial electrograms in human persistent atrial fibrillation," *Medical & biological engineering & computing*, vol. 56, no. 1, pp. 71–83, 2018.

[7] D. Satoh, Y. Takano, R. Sudo, and T. Mochida, "Reduction of communication demand under disaster congestion using control to change human communication behavior without direct restriction," *Computer Networks*, vol. 134, pp. 105–115, 2018.

[8] L. Z. Gruber, A. Haruvi, R. Basri, and M. Irani, "Perceptual dominance in brief presentations of mixed images: Human perception vs. deep neural networks," *Frontiers in Computational Neuroscience*, vol. 12, p. 57, 2018.

[9] P. Loncomilla, J. Ruiz-del Solar, and L. Martínez, "Object recognition using local invariant features for robotic applications: A survey," *Pattern Recognition*, vol. 60, pp. 499–514, 2016.

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[11] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European conference on computer vision*. Springer, 2008, pp. 650–663.

[12] J. Zhang, Z. Zhu, W. Zou, P. Li, Y. Li, H. Su, and G. Huang, "Fastpose: Towards real-time pose estimation and tracking via scale-normalized multi-task networks," *arXiv preprint arXiv:1908.05593*, 2019.

[13] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[14] R. Mask, "Kaiming he, georgia gkioxari, piotr dollr, and ross girshick," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[15] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2334–2343.

[16] U. Iqbal, A. Milan, and J. Gall, "Posetrack: Joint multi-person pose estimation and tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2011–2020.

[17] G. Chartrand, *Introduction to graph theory*. Tata McGraw-Hill Education, 2006.

[18] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[19] Z. Zhang, K. Huang, T. Tan, and L. Wang, "Trajectory series analysis based event rule induction for visual surveillance," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.