An assault detection system based on human Pose Tracking for video surveillance

Pedro G. S. do Couto Soares*, Arnaldo V. Barros da Silva*, Luis F. Alves Pereira*
*Universidade Federal Rural de Pernambuco, Unidade Acadêmica de Garanhuns,
Pernambuco, Garanhuns 55292-270
Email: pscoutosoares@gmail.com

***Abstract*—The development of new technologies for video surveillance and automatic violence detection can bring more security to our daily lives. Solutions previously published in the state-of-the-art had presented techniques to detect violence at movie scenes, sports matches, or crowds. In this work, we propose a novel system architecture based on human Pose Track for detecting evidence of assaults in real-world videos from closed-circuit television (CCTV) of Brazilian lottery agencies. The results showed that our method can identify individuals with hands up and lying down with accuracy rates up to 85%. We believe that the detection of potentially risky situations in real-time is a crucial tool in the fighting against crime.**

I. Introduction

Video action recognition is a hot topic in computer vision [1]–[3]. Within the wide range of applications that arises from those methods, automatic violence detection [4]–[8] has high potential of impact in public and private security due to *(i)* the high number of surveillance cameras deployed everywhere, and *(ii)* the impracticability of employing people for monitoring such videos in real-time.

Security is, indeed, a serious concern worldwide. However, in emerging countries, the crime rates are increasing to extremely high levels [9], [10]. In Brazil, for instance, lottery agencies are robbed very often [11]–[14] due to the high movement of money in those places. Thus, this work proposes an action recognition system that processes images of closed-circuit television (CCTV) of Brazilian lottery agencies and generates alerts in real-time when a potential assault is in progress. Related works on violence detection [4]–[8], [15]–[18] show solutions applied to crowds, movie scenes, and sports matches in a context very different than the proposed here.

We assume that the probability of catching thieves is higher as soon the security agents receive alerts about potential assaults in progress. Nowadays, the crime is only reported after its end, and the CCTV images are used for starting a lengthy investigation.

The method proposed in this work is based on human Pose Tracking [19] that estimates the position of key-points at individuals body - such as the left knee, right knee, left elbow, right elbow - along with the frames that compose the video. Such data is encoded to represent human actions and then classified. The identification of people with hands up or laying down is potentially related to an alarming situation.

To the best of our knowledge, no previous publication from the state-of-the-art *(i)* shows the performance of violence detection algorithms at real-world videos of assaults, or *(ii)* present a solution for this problem based on Pose Tracking.

The Pose Estimation method applied in the proposed architecture employes modern Deep Learning techniques to identify persons and their body parts. Such method allows a highly accurate and real-time identification.

This paper is organized as follows: Section II presents the architecture of the proposed method for assault detection; Section III describes the dataset used in the experiments and also shows the classification rates obtained by the proposed method; finally, Section IV concludes the paper.

II. Proposed Method

The architecture of the proposed system is presented in Fig. 1. It shows that for each new video given as input, a set of K frames $\{h_k\}$ where $k \in \{-\frac{K}{2}, \cdots, 0, \cdots, \frac{K}{2} - 1\}$ are selected to be processed together according to the following steps:

1) **pose estimation**: to identify 17 interest points at each human body present in the frames h_k by using the framework for regional multi-person estimation proposed by Fang et al. [20];
2) **pose tracking**: to track the interest points identified along the K frames using the Pose Flow technique proposed by Xiu et al. [19];
3) **feature extraction**: to generate a single feature vector that encodes the pose tracking data acquired in the actual K frames;
4) **classification**: to classify the human behavior encoded by the feature vector previously generated.

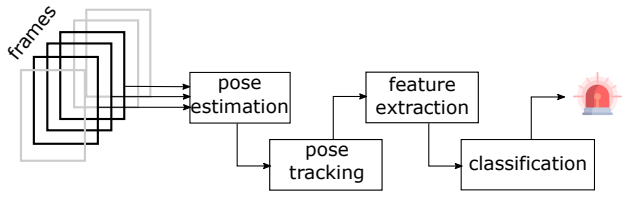

Fig. 1. Architecture of the system proposed for assault detection.

In the **feature extraction** module, we proposed a template of distances to store the relative distances between each of the

17 interest points previously identified at the **pose estimation** module. Such template is encoded into a vector $\mathbf{c}$ of $\binom{17}{2} = 136$ dimensions, as shown in Fig. 2. As a result, K vectors $\mathbf{c}$ are created to represent the pose of a human along the set of frames $\{h_k\}$. Then, a final representation $\mathbf{r}$ of the human activity along the K frames is composed of the concatenation between the vector of averages $avg(\mathbf{c_k})$ and the vector of standard deviations $std(\mathbf{c_k})$, *i.e.*, $\mathbf{r} = [avg(\mathbf{c_k}); std(\mathbf{c_k})]^T$.

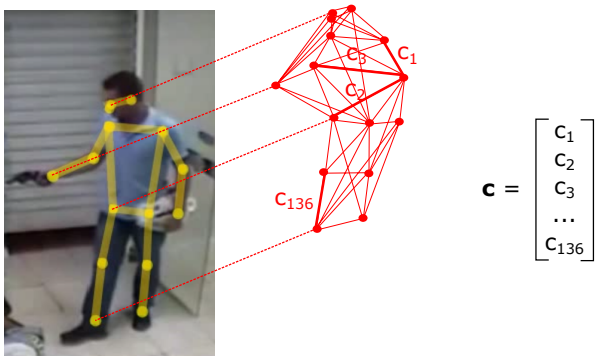

Fig. 2. Template of distances encoded into a vector $\mathbf{c}$ during the feature extraction phase.

In the **classification** module, the vector $\mathbf{r}$ is given as input of different SVMs, each one is able to identify distinct human activities. In this work, two SVMs with an RBF kernel are evaluated detect individuals with *hands up* or *lying down*, *i.e.*, movements that indicate potentially alarming situations. SVM classifier was chosen since it was widely used in many others systems for violence detection [4]–[7] and showed high accuracy.

III. Experiments

Experiments were conducted to evaluate the accuracy of the classification of human actions that indicate potential risk of assaults. In the next subsections, the database created for this work is presented, and the classification results are shown.

A. Database

Due to the lack of datasets available in the literature with real-world videos of assaults at indoors establishments, a new benchmark for violence detection was created.

The dataset created is composed of $47,280$ frames extracted from CCTV records downloaded from YouTube. The videos contain scenes of assaults and normal days at many Brazilian lottery agencies. Fig. 3 illustrates samples from our dataset.

The movements of each individual in the frames were manually labeled as *hands up* or *lying down* when they occur. Furthermore, labels *assaults* and *non-assaults* were assigned to each complete video in the dataset.

B. Classification results

The classification results of each SVM are presented in details at the confusion matrices of Tables I and II. The performance measures extracted from such data is presented in Table III, and the performance of both classifiers at different operation points are presented in the ROC curves of Fig. 4.

Fig. 3. Samples of the videos in the dataset created.

The reader may check that both classifiers achieved recall rates up to 89%. It means that only a small number of *hands up* or *lying down* movements are not detected by the proposed solution. The precision rates are lower than the recall rates. It means that some movements detected by the system are false alarms.

Those performance measures indicate that the proposed system is a useful tool for alerting security agents about potential risky situations. Since agents cannot be watching all the cameras every time, they will check only the positive classifications to prevent false alarms and - when assaults are confirmed - they can move quickly to the lottery agency.

TABLE I
CONFUSION MATRIX THAT SHOWS THE CLASSIFICATION RESULTS OF INDIVIDUALS WITH HANDS UP

		Prediction outcome: False	Prediction outcome: True	total
Actual value	False	35	7	42
	True	5	37	42
	total	40	44	

IV. Conclusion

We presented a technique for assault detection based on human Pose Tracking. Experiments conducted using a database of real-world CCTV videos of Brazilian lottery agencies

TABLE II
CONFUSION MATRIX THAT SHOWS THE CLASSIFICATION RESULTS OF INDIVIDUALS LYING DOWN

		Prediction outcome: False	Prediction outcome: True	total
Actual value	False	59	29	88
	True	9	79	88
	total	68	108	

TABLE III
PERFORMANCE METRICS OF CLASSIFICATIONS

	Hands up	Lying down
Precision	84%	73%
Recall	88%	89%
Accuracy	85%	78%

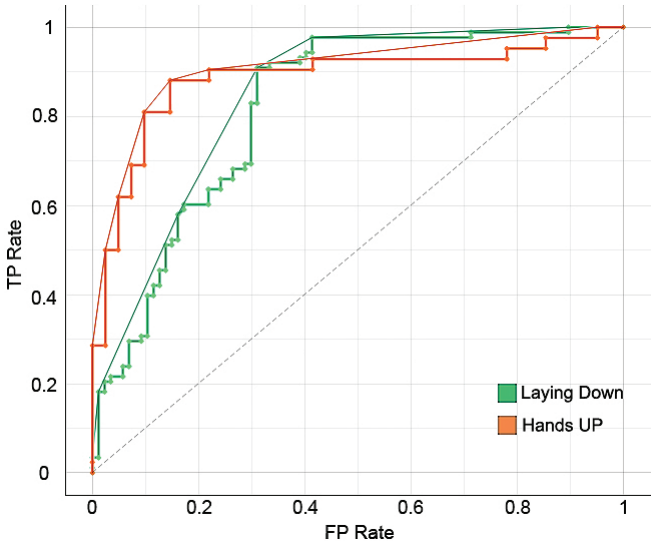

Fig. 4. ROC curves for classification of human activities in videos.

showed that our method obtained accuracy rates up to 85% in the detection of potentially alarming situations with people with hands up or lying down.

Future works should involve the identification of other human actions such as *draw a gun*, *run*, and others. Finally, a complete solution of video understanding for surveillance of indoor establishments can be developed and launched in the market.

REFERENCES

[1] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[2] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[3] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in human action recognition: A survey," *arXiv preprint arXiv:1501.05964*, 2015.

[4] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image and vision computing*, vol. 48, pp. 37–41, 2016.

[5] P. C. Ribeiro, R. Audigier, and Q. C. Pham, "Rimoc, a feature to discriminate unstructured motions: Application to violence detection for video-surveillance," *Computer vision and image understanding*, vol. 144, pp. 121–143, 2016.

[6] T. Senst, V. Eiselein, A. Kuhn, and T. Sikora, "Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2945–2956, 2017.

[7] P. Zhou, Q. Ding, H. Luo, and X. Hou, "Violence detection in surveillance video using low-level features," *PLoS one*, vol. 13, no. 10, p. e0203668, 2018.

[8] A. Hanson, K. Pnvr, S. Krishnagopal, and L. Davis, "Bidirectional convolutional lstm for the detection of violence in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[9] A. Erickson, "Latin america is the worlds most violent region. a new report investigates why," April 2018, the Washington Post, [posted 25-April-2018]. [Online]. Available: https://wapo.st/2Gq55xN

[10] D. Luhnow, "Latin america is the murder capital of the world," september 2018, the Wall Street Journal, [posted 20-September-2018]. [Online]. Available: https://www.wsj.com/articles/400-murders-a-day-the-crisis-of-latin-america-1537455390

[11] "Segurana morto em tentativa de assalto lotrica na terra firme," April 2019, g1 - O portal de notcias da Globo, [posted 30-April-2019]. [Online]. Available: https://g1.globo.com/pa/para/noticia/2019/04/30/seguranca-e-morto-em-tentativa-de-assalto-a-loterica-na-terra-firme.ghtml

[12] "Vdeo mostra assaltos a casas lotricas no interior de alagoas," March 2019, g1 - O portal de notcias da Globo, [posted 18-March-2019]. [Online]. Available: https://g1.globo.com/al/alagoas/noticia/2019/03/18/video-mostra-assaltos-a-casas-lotericas-no-interior-de-alagoas.ghtml

[13] R. Leal, "Empresrio assaltado em lotrica diz que criminoso se passou por cliente para roubar dinheiro," June 2019, g1 - O portal de notcias da Globo, [posted 19-June-2019]. [Online]. Available: https://glo.bo/30Y9FLv

[14] C. Costa, "Vdeo mostra ao de criminosos em assalto a lotrica em parnaba," May 2019, g1 - O portal de notcias da Globo, [posted 30-May-2019]. [Online]. Available: https://g1.globo.com/pi/piaui/noticia/2019/05/30/video-mostra-acao-de-criminosos-em-assalto-a-loterica-em-parnaiba.ghtml

[15] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition*. IEEE Computer Society, 2009, pp. 2929–2936.

[16] S. Blunsden and R. Fisher, "The behave video dataset: ground truthed video for multi-person behavior classification," *Annals of the BMVA*, vol. 4, no. 1-12, p. 4, 2010.

[17] E. B. Nievas, O. D. Suarez, G. B. García, and R. Sukthankar, "Violence detection in video using computer vision techniques," in *International conference on Computer analysis of images and patterns*. Springer, 2011, pp. 332–339.

[18] K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human action classes from videos in the wild," *Center for Research in Computer Vision*, 2012.

[19] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," *arXiv preprint arXiv:1802.00977*, 2018.

[20] 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.