Hands Off: A Handshake Interaction Detection and Localization Model for COVID-19 Threat Control

A. S. Jameel Hassan[†], Suren Sritharan[‡], Gihan Jayatilaka[†],
Roshan I. Godaliyadda[†], Parakrama B. Ekanayake[†], Vijitha Herath[†], Janaka B. Ekanayake[†]
[†]*Department of Electrical and Electronic Engineering, University of Peradeniya, Sri Lanka*
[‡]*School of Computing and IT, Sri Lanka Technological Campus, Sri Lanka*
{jameel.hassan.2014, suren.sri, gihanjayatilaka}@eng.pdn.ac.lk,
{roshangodd, mpb.ekanayake}@ee.pdn.ac.lk, {vijitha, ekanayakej}@eng.pdn.ac.lk,

***Abstract*— The COVID-19 outbreak has affected millions of people across the globe and is continuing to spread at a drastic scale. Out of the numerous steps taken to control the spread of the virus, social distancing has been a crucial and effective practice. However, recent reports of social distancing violations suggest the need for non-intrusive detection techniques to ensure safety in public spaces. In this paper, a real-time detection model is proposed to identify handshake interactions in a range of realistic scenarios with multiple people in the scene and also detect multiple interactions in a single frame. The efficacy of the proposed model was evaluated across two different datasets on more than 3200 frames, thus enabling a robust localization model in different environments. The proposed model is the first dyadic interaction localizer in a multi-person setting, which enables it to be used in public spaces to identify handshake interactions and thereby identify and mitigate COVID-19 transmission.**

***Index Terms*—COVID-19, deep learning, human-human interactions, dyadic interaction localization**

I. Introduction

The novel COVID-19 virus is one of the biggest threat to global health since the Spanish flu in 1918. As of July 2021 nearly 196 million people have been infected and more than 4 million people have succumbed to death due to the virus [1]. Vaccination has been identified as the most effective measure by the World Health Organization (WHO) to curtail the spread of the virus [2]. However, complete vaccination of the entire global population has not been possible due to varying production and logistic issues. Therefore, the key measure taken for the curtailment of the spread of COVID-19 has been social distancing.

Social distancing has been found to be a promising approach towards mitigating the virus spread [3]. Nevertheless, humans as a social species, tend to deviate from such constrained behavior [4] for prolonged periods of time. Thus, it is crucial to identify such breach of social distancing protocols in order to ensure the safety of the society. Importantly, human-human interactions need to be ensured minimal as it is the most severe form of breach which are also the easiest to avoid. Moreover, a simple greeting is the often the initial breach of social distancing. Therefore, identifying and localizing such interactions such as from a CCTV footage will enable to create a framework to prevent such breach of social distancing measures.

Identification of human interactions often referred to as dyadic interactions (interactions between two people) has been explored in the action recognition domain. Action recognition has moved from an object detection/tracking problem [5], [6] in to a multi-class classification problem. Dyadic interaction detection has spawned from human action recognition in computer vision literature. Most of action recognition has focused on a single person in the frame performing a specific action such as running, walking, jumping etc [7]. Recently works have focused on behavior/activity recognition of multiple people in the frame, such as in a game [8].

The use of limb positions to identify interactions was presented in [9]. The idea stemmed from the concept that each interaction presented unique limb positioning. As a next step, considering the gross body movement and proximity measures was done in [10]. This is done in a multi-step manner where the person localization is used for the interaction identification. This poses a drawback in interaction localization as the error in the first stage of person localization can extend to the next stage. This has been improved by considering this a multiple instance learning problem by [11] since not all frames in an interaction are considered informative.

Bag of visual words method has also been used to identify body movements. Local features from this are pooled and a mapping is generated from this to interactions in [12]. Part-based models such as deformable part model (DPM) [13] has been proven extremely effective in people and body part detection and localization prior to neural networks. The use of interaction specific DPMs to identify people in specific poses is done in [14]. An extension of this work using spatio-temporal DPMs to localize dyadic interactions has been presented in [15]. This has been one of the few works that localizes the interaction itself instead of the actors.

The advent of neural networks has drastically overtaken DPM techniques in detection problems. The YOLO network [16] is a highly robust neural network capable of detecting 80 classes in real-time (78 FPS). In [17] a human activity recognition model has been formulated using the YOLO network on the LIRIS dataset [18]. Most notably, this model can perform the localization in real-time which is crucial depending on the

This paper has been accepted for ICIIS 2021

need. A recurrent neural network (RNN) based spatio-temporal attention mechanism for human interaction recognition is performed in [19]. This model incorporates attention to the hands of the body to identify the interaction. However, one of the main drawbacks of this and other methods is the absence of real-time detection. The above cited works except [14], [15] in dyadic interaction detection consider a video feed/frame and classify it to the given class of interaction or identify the actors where the localization of the interaction is not considered. This localization too is performed only with two persons in the frame.

The major contributions of this paper are as follows. In this paper, the first human interaction localization model in a multi-person setting is proposed. A handshake interaction localization model in real-time to mitigate the threat for COVID-19 is presented using computer vision in a non-intrusive technique. This ensures a scalable, robust model that can be used in public spaces and work environments to mitigate the spread of COVID-19.

II. Proposed solution

A convolutional neural network (CNN) based model is proposed to identify and localize handshake interactions in a multi-person setting for wall mounted CCTV video footage. The model architecture used is the YOLO network with training and testing performed using a novel dataset and the UT-interaction (UTI) [20], [21] dataset.

A. YOLO network

The YOLO network is a state-of-the-art (SOTA) CNN in object detection. It was the pioneering work in creating a one-stage detection network for the object detection task. The key change in YOLOv3 [22] was the approach to divide the image into grids (such as 13×13), and then predict a fixed number (such as 3) of bounding boxes for each grid cell. The bounding box is predicted with the relevant class and object confidence score. The architecture of the YOLOv3 network is shown in Fig.1.

The YOLOv3 architecture makes prediction at three stages in the neural network depth as seen in Fig.1. This enables detection of objects of all sizes, which was the main drawback in previous versions. The first stage detector outputting a 13×13 grid is better at predicting larger images, while the 26×26 grid predicts medium sized images and the 52×52 grid prediction in stage three is best at predicting small images. The image input (resized to 416×416 passes through the convolutional layers to output a tensor of shape $h\times h\times18$. Here h is the number of grid cells along one axis and 18 corresponds to $3\times(5+1)$, where 3 is the number of bounding boxes predicted in one grid cell, 5 is the number of bounding box attributes and 1 is the number of classes. The bounding box attributes are the coordinates of the four vertices and the objectness score.

B. Model Training

In order to train the YOLO network for handshake interaction detections, a suitable dataset is required. There are few

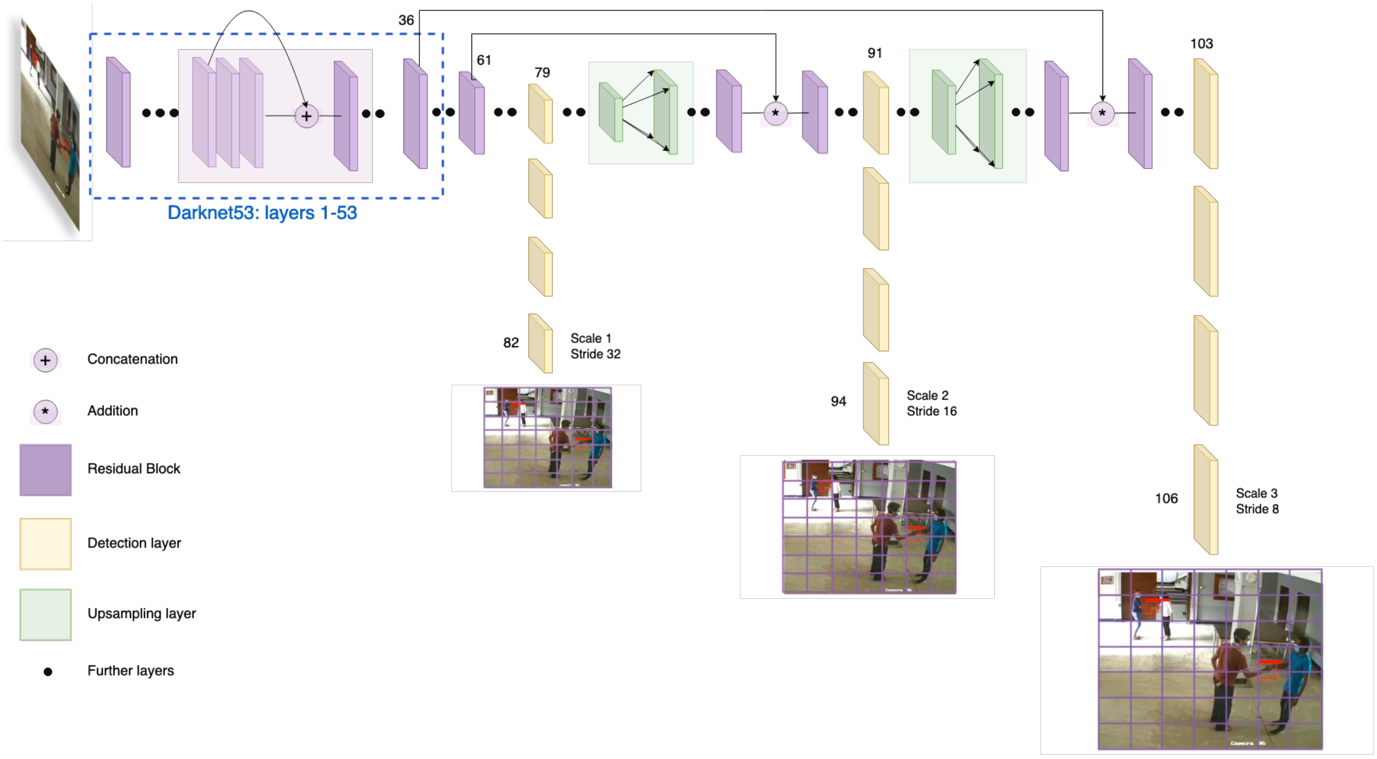
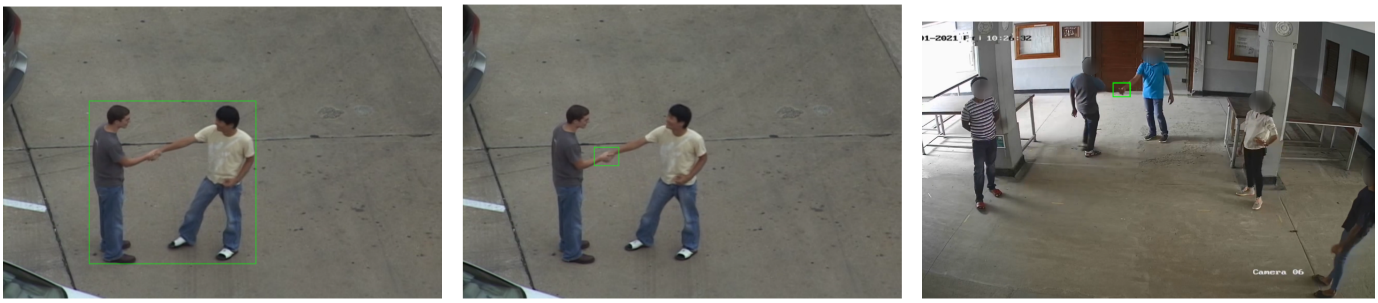


Fig. 1: YOLOv3 architecture.



(a) Original ground truth for UTI.

(b) Created ground truth for UTI.

(c) Ground truth of Shakes.

Fig. 2: Dataset ground truth annotations.

datasets in the action recognition domain for computer vision. However, the handshake interactions in these datasets are minimal and even then, the ground truth for such datasets are not for the localization problem but for actors identification. Furthermore, existing datasets for dyadic interactions have only two people in the frame. Since our motivation is to identify interactions to combat COVID-19, a video footage with multiple people in the frame, where dyadic interactions occur is necessary. Therefore, a dataset rich in context to tackle the problem of human interaction identification in a multi-person setting was created. The existing UTI dataset was also used in the framework by relabelling the handshake interactions for the localization problem.

C. Datasets

The existing UTI dataset was re-labelled by marking the interactions in each frame. Fig.2 shows the original ground truth and the created ground truth data for the UTI dataset.

Due to the scarcity of handshake interactions, a new dataset was created in the university premises using wall mounted CCTVs. This consisted of 10 videos each spanning nearly 1500 frames. We refer to this dataset as the ”Shakes dataset”. This consists a multi-person setting and also multiple interactions in the same instance in many frames. A sample frame is shown in Fig.2c.

D. Training Using Transfer Learning

In order to train the YOLO network for the handshake interactions, the darknet53 (highlighted by a cyan dotted rectangle) Fig.1, referred to as the YOLO backbone was initialized with weights obtained by training on the Imagenet dataset [23]. Then, 3000 images of hands from the open images database [24] were used for training the YOLO network as the first stage, since a larger distribution of images was available here. Using the weights of the network from this training phase, the handshake images from the Shakes and UTI datasets were trained. The transfer learning approach was used as the handshake interactions were from a smaller distribution. Out of 20 videos, 17 videos from the UTI dataset and 5 out of the 10 videos from the Shakes dataset were used. While the number of videos from the UTI dataset is higher, the number of frames were maintained approximately equal.

III. Results and Discussion

The model was evaluated using both the aforementioned datasets. The Average precision (AP) and the Mean average precision was used as the evaluation metric, as prominent object detection competitions such as PASCAL VOC challenge [25], COCO detection challenge [26] and the Google Open Images dataset [24] competition use these metrics as key parameters in evaluating the detector performance.

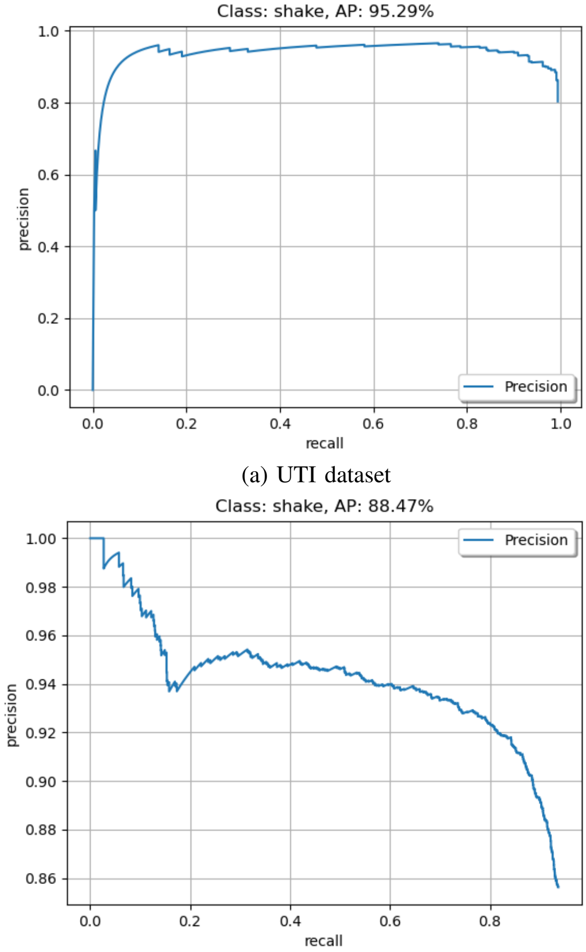
The Average precision (AP) is the precision value averaged across varying recall values between 0 and 1. This is computed using area under the curve (AUC) of the precision vs recall curve, plotted as a function of the confidence threshold of detection with a constant intersection over union (IoU) for the bounding box threshold [27]. This IoU threshold is usually maintained at 0.5 in object detection tasks.

The performance of the model in detecting handshake interactions was evaluated on the UTI and Shakes dataset separately and is tabulated in Table I. 3 videos containing 418 frames from the UTI dataset and 5 videos with 2786 frames from the Shakes dataset were considered for this purpose. The AP value for the UTI dataset was 95.29% and for the Shakes dataset was 88.47%. The precision vs recall curves for the UTI dataset and the Shakes dataset are shown in Fig.3.

TABLE I: Performance metrics of handshake detection

Dataset	**AP/%**
UT-interaction	95.29
Shakes	88.47

Few frames of detection and localization of handshake interactions from the UTI and the Shakes dataset are shown in Fig.4. It can be observed that the neural network can identify more than just a single handshake interaction in the frame. It is also able to identify interactions even at different scales as seen in Fig.4c. A more realistic setting such as interactions in a busy public space is shown in Fig.5. The neural network performs well to even detect such interactions such as that might occur in an office corridor or a busy public place. The neural network was also tested for very rare cases by considering hand occlusion cases. Fig.5b shows instances of



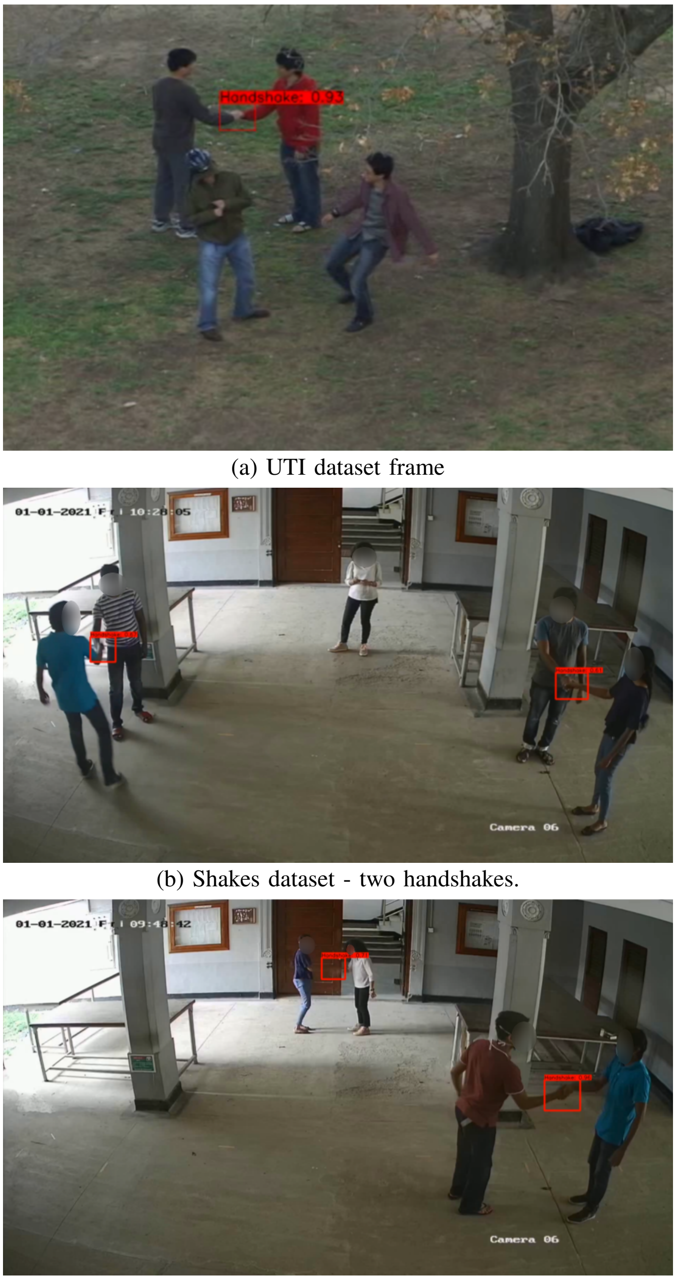
(a) UTI dataset

(b) Shakes dataset

Fig. 3: Precision vs Recall curves for handshake localizer for UTI and Shakes dataset.

such occlusions intentionally mimicking a handshake which the neural network avoids detecting.

Finally, the false positives of the neural network model in handshake interaction localization were analyzed. The false positives can be observed in Fig.6. It can be observed that most errors occur during occlusion or in instances where the hand positions are similar to those during handshakes, ie: an outstretched hand and palm. Furthermore Fig.6f depicts an instance where one handshake is identified whilst the other is not.



(a) UTI dataset frame

(b) Shakes dataset - two handshakes.

(c) Handshakes at different distances from the camera.

Fig. 4: Handshake interaction detection localizations.

Conclusion

Lack of human adherence to social distancing protocols is notably increasing thereby compounding the spread of COVID-19. This demands a scrutinized monitoring of human interactions in public spaces to identify and mitigate such violations of social distancing measures. In this paper, we present a neural network model to identify handshake interactions in realistic scenarios from CCTV footage in a multi-person setting. The neural network performance is validated by comparing its localization in 2 different datasets. The ability to detect and localize interactions in real-world settings and the detection of multiple interactions in a single frame affirm the robust nature of the model. The deployment of this model will enable us to identify social distancing violations in real-time and thereby create a framework to reduce such violations and mitigate the adverse impacts of COVID-19. As a deployable system, the model could be further improved to localize more challenging interactions such as hugs and kisses to combat the pandemic.

Acknowledgements

This work was funded by Lewis Power, Singapore and International Development Research Centre (IDRC), Canada. The authors are thankful to the volunteers for their support in creating the Shakes dataset.

References

[1] Worldometer, https://www.worldometers.info/coronavirus/, 2021, last accessed on 2021-07-20.

[2] W. H. Organization, "WHO - Covid-19 vaccines," https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines, 2021, last accessed on 2021-07-20.

[3] M. Qian and J. Jiang, "Covid-19 and social distancing," *Journal of Public Health*, pp. 1–3, 2020.

[4] A. Venkatesh and S. Edirappuli, "Social distancing in covid-19: what are the mental health implications?" *Bmj*, vol. 369, 2020.

[5] W. Fernando, H. Herath, P. Perera, M. Ekanayake, G. Godaliyadda, and J. Wijayakulasooriya, "Object identification, enhancement and tracking under dynamic background conditions," in *7th International Conference on Information and Automation for Sustainability*. IEEE, 2014, pp. 1–6.

[6] R. Rupasinghe, S. Senanayake, D. Padmasiri, M. Ekanayake, G. Godaliyadda, and J. Wijayakulasooriya, "Modes of clustering for motion pattern analysis in video surveillance," in *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*. IEEE, 2016, pp. 1–6.

[7] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7912–7921.

[8] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei, "Detecting events and key actors in multi-person videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3043–3053.

[9] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *European conference on computer vision*. Springer, 2010, pp. 168–181.

[10] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in tv shows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2441–2453, 2012.

[11] F. Sener and N. Ikizler-Cinbis, "Two-person interaction recognition via spatial multiple instance embedding," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 63–73, 2015.

[12] M. J. Marin-Jimenez, E. Yeguas, and N. P. De La Blanca, "Exploring stip-based models for recognizing human interactions in tv videos," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1819–1828, 2013.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[14] C. Van Gemeren, R. Poppe, and R. C. Veltkamp, "Spatio-temporal detection of fine-grained dyadic human interactions," in *International Workshop on Human Behavior Understanding*. Springer, 2016, pp. 116–133.

[15] ——, "Hands-on: deformable pose and motion models for spatiotemporal localization of fine-grained dyadic interactions," *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, pp. 1–16, 2018.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[17] S. Shinde, A. Kothari, and V. Gupta, "Yolo based human action recognition and localization," *Procedia computer science*, vol. 133, pp. 831–838, 2018.

[18] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.-E. Bichot *et al.*, "Evaluation of video activity localizations integrating quality and quantity measurements," *Computer Vision and Image Understanding*, vol. 127, pp. 14–30, 2014.

[19] F. Baradel, C. Wolf, and J. Mille, "Human action recognition: Pose-based attention draws focus to hands," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 604–613.

[20] M. S. Ryoo and J. K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)," http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.

[21] ——, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[22] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

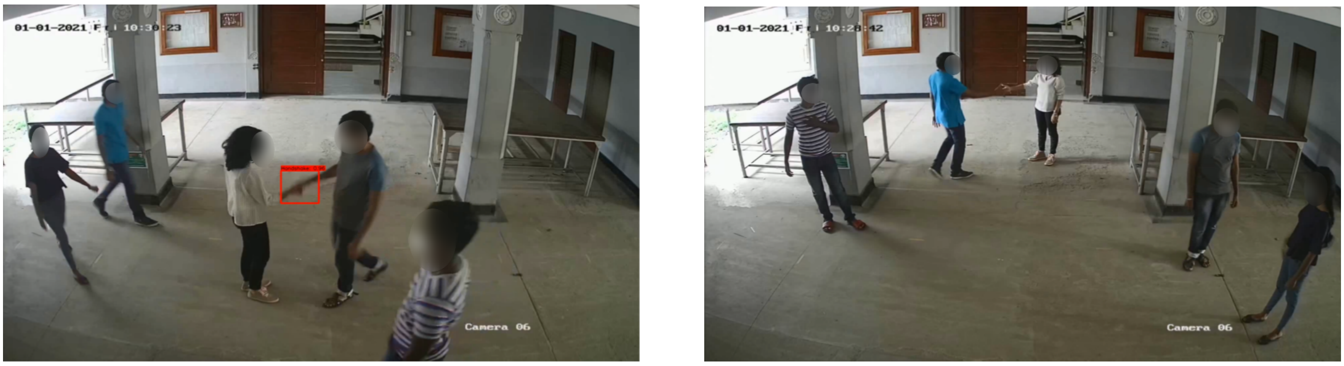
[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[24] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *arXiv preprint arXiv:1811.00982*, 2018.

[25] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

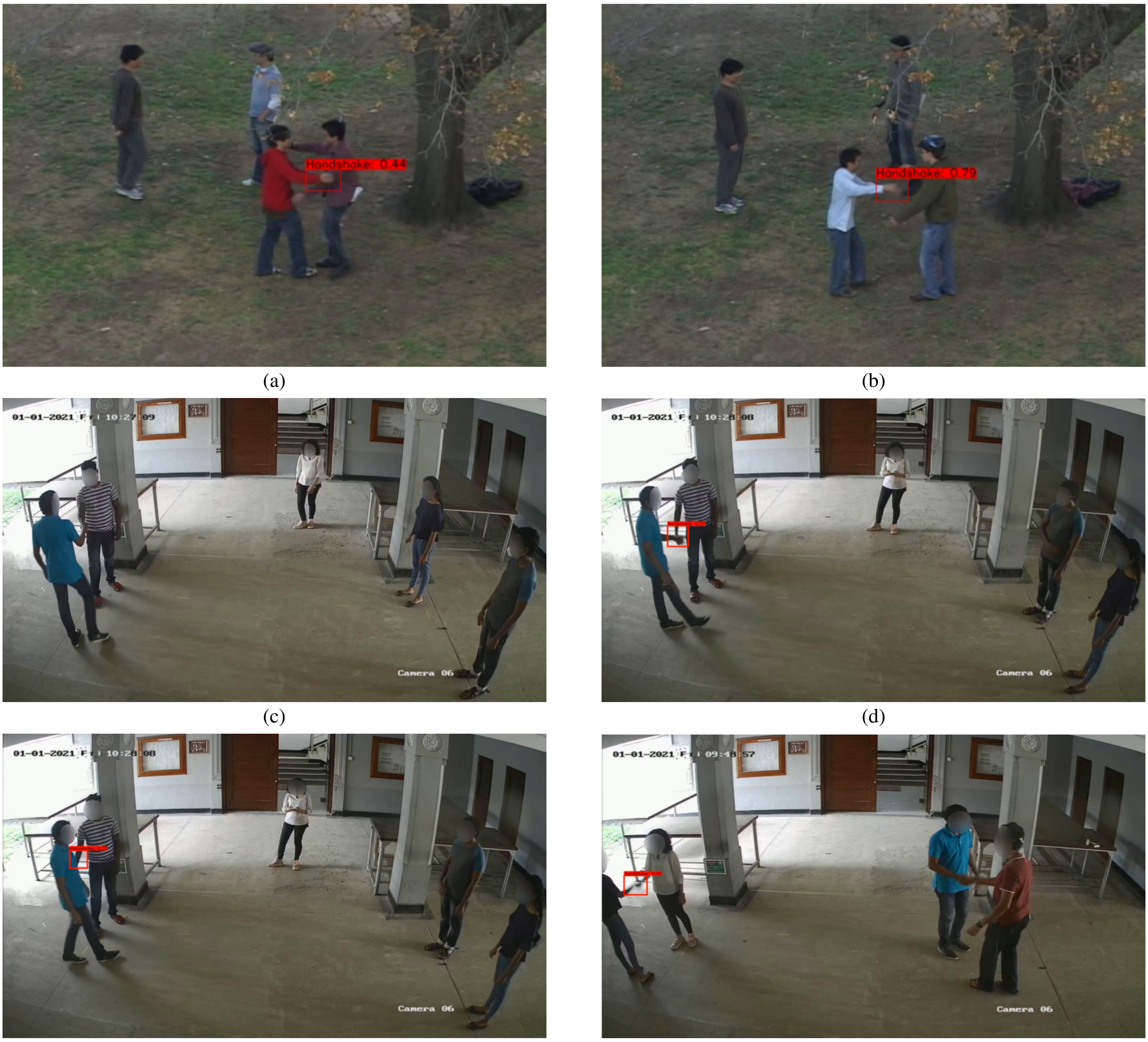
[27] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020, pp. 237–242.



(a) Handshake in corridor.

(b) Fake handshake by occlusion.

Fig. 5: Detection localizations in busy settings and occlusion cases.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 6: Localization false positives and false negatives in the model.