ICIC Express Letters
Part B: Applications
Volume **13**, Number **11**, November **2022**
 ISSN 2185-2766
pp. 1177–1187

COMBINING SVM AND HUMAN-POSE FOR A VISION-BASED FALL DETECTION

Atchara Namburi* and Thapani Hengsanankun

Department of Computer and Information Science
Faculty of Science and Engineering
Kasetsart University
59/5 Moo1 Chiangkrua, Muang, Sakon Nakhon 47000, Thailand
thapranee.h@ku.th
*Corresponding author: csnarn@ku.ac.th

Received March 2022; accepted May 2022

Abstract. *In this paper, we propose a method for detecting a human fall in colour video sequences. We combine a Support Vector Machine (SVM) with the Open-Pose method to classify parts of individual frames as human falling or not falling. We use Open-Pose to detect the 25 human body keypoints. We select two keypoints: neck and mid abdomen, to represent a human upper body. We subsequently draw a line on those keypoints to obtain the angle against the Y-axis. The angle is used as a feature for SVM to train. Finally, we evaluated our work by using five datasets. We demonstrate good results on a number of videos, comparing the detection and identification produced by this method with the manually extracted truth. We obtain 89.66% average classification accuracy and 90.78% and 88.27% of the sensitivity and specificity, respectively.*
Keywords: Fall detection, SVM, YOLOv4, Open-Pose, deepSORT

1. **Introduction.** Fall is one of the most dangerous accidents that threaten health in daily life. Thailand's public health statistics report that approximately 30% of older people fall once or more annually. About 1.67% of those injured die each year in Thailand [27]. In addition to that, 20% of those who fall are injured [27]. More than that, the World Health Organization (WHO) [20] reveals statistics about key facts about fall causes. They stated that approximately 684,000 individuals pass away from falls globally. Falls can lead to different injuries and be even more severe for seniors, such as head injury, hip fracture, and heart attack if immediate help is unavailable [25]. It can lead cause of death ultimately in the absence of any support. In other words, a person who falls will probably have no progressive severity if the medical treatment arrives in time. We notice that falls among the elderly are a big problem; it is vital to look into a system that can detect, and alarm falls immediately to reduce the duration of harm. The earlier the fall is detected, the lower the morbidity and mortality rate [11, 30].

In the last decade, many surveillance systems have been developed for fall detection and alert. There are lots of work presented in the state of the art for fall detection. Several proposed works aim to decrease injuries for the senior person and prevent severe injuries for people who fall. Different materials have been used to detect falls. However, fall detection can be categorized into two main categories: hardware-reliant-based and video-based. The hardware-reliant-based utilized sensors to generate data and then detect the person's posture and activities [29], such as Inertial Measurement Units (IMU) sensor. In addition, some works employed the ambient such as vibration, audio, and videos for detection. The hardware-reliant-based benefits the accuracy provided by devices. Sensors can provide high sensitivity and specificity and also give high accuracy. However, some

DOI: 10.24507/icicelb.13.11.1177

drawbacks for the elderly such as a senior person needing to wear it may raise a wounded or discomfort caused by the sensor [16] and forgetfulness. In addition, the ambient sensor is relatively expensive and unaffordable for middle-income families [5].

In contrast, the method based on computer vision offers advantages. It is a low-cost and less intrusive system, and no physical contact is needed because the older person does not need to wear it. The system becomes part of their dwelling when it is installed and set up. However, the privacy concerns come along with using videos to monitor a fall event [31]. The vision-based technique focuses on the video sequences to distinguish fall events by using the deformation of human shape (the change of the body shape is calculated), and the bounding box is used to surround the object in the videos. Many works employ a bounding box to detect falls, calculating the aspect ratio or the angle of the box to identify the different activities. However, using only a bounding box is insufficient to measure the posture deformation accurately. Therefore, more information is necessary to differentiate posture.

Some works employed the human-pose estimation to detect human falls. For example, [21] proposed the skeleton estimation method, which is used to extract 25 joint human points. Some system uses background subtraction [4, 24, 28] to distinguish the person against the background and notify if a person is fallen. However, there is a report that the "tucked" fall and occlusion could not be detected. Another shortcoming was that the system only detected just one person. The occlusion problem was addressed in [8] using multi-cameras to help the occluded part of a person in some camera frames; however, not many results were presented. There is some approach to adopting the tracking system to recognize the fall event [7, 12, 15, 18, 24]. In a stationary area such as the floor, they recognize that they utilize the relational data among the area of moving objects. The tracking strategy is performed with the initial conditions, which can suffer when not met. The Machine Learning (ML) technique has been gaining popularity over the last decade. Many researchers adopt ML to differentiate fall event from other everyday activities [3, 9, 19].

Even if a wide range of studies has shown promising results, there still is a vast space to improve fall detection. In this paper, we propose a video-based fall detection approach that can improve the performance of vision-based fall detection for the elderly. We are combining the well-known classifier SVM and augmented features of human pose data from Open-Pose. Firstly, we use Open-Pose to detect the 25 human body key points. We then pick two key points: neck and mid-abdomen, representing a human upper body. We draw a line on those key points to obtain the angle against the Y-axis. The angle is used as a feature for SVM to train.

The rest of this paper is organized as follows. Section 2 proposes our approach in detail. Section 3 indicates results, and we conclude our work and provide a prospect for future work in Section 4.

2. **Proposed Method.**

This section proposes our work, the fall detection based on SVM that employs the human key point, resulting from human pose estimations, as to the input feature vector. Figure 1 illustrates the overview of our proposed fall detection. Firstly, we used the Open-Pose to extract the 25 human key points. We then pick two essential key points: neck and mid-abdomen, representing the upper human body. Next, we calculate the body orientation angle against the Y-axis used as the SVM input feature vector. In the testing step, we work parallel YOLOv4 with deepSORT and Open-Pose to track multiple persons. We employ Open-Post to extract the human pose skeleton. We then calculate the angle of the human upper body against the Y-axis as the input feature vector for SVM. There are 346 colour videos from five different datasets used for our work, 214 videos for training, and 132 videos for testing. Figure 2 shows examples of a fall and a non-fall of the orientation angle and the ground truth.

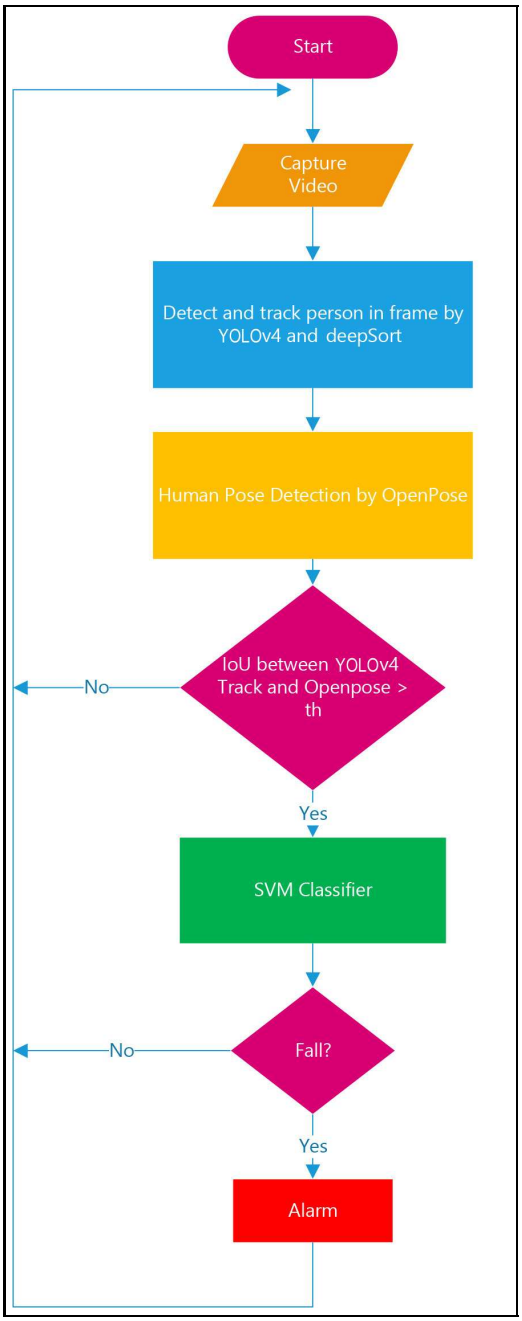

FIGURE 1. The overview of our fall detection system

2.1. **Human pose detection.** Open-Pose [21] is one of the popular bottom-up approaches to estimating human pose. It initially detects every person's key points in the image, then assigning them to distinct individuals. It jointly detects the human body, hand, facial, and foot. This paper uses the OpenPose body 25 human description, which provides 25 human body key points in 2D to describe human pose as shown in Figure 4.

The definition of fall in the Cambridge online dictionary [2] is "to suddenly go down onto the ground or towards the ground without intending to or by accident". In addition, the Kellogg International Work Group on the Prevention of Falls by the elderly [1] defines the fall as "unintentionally coming to the ground, or some lower level not as a consequence of sustaining a violent blow, loss of consciousness, sudden onset of paralysis as in stroke or an epileptic seizure". The definition has been used in many research studies, as it is used in this paper to clarify the body orientation during falls. Therefore, we have picked two key points to represent the human upper body: neck and mid-abdomen, as shown in

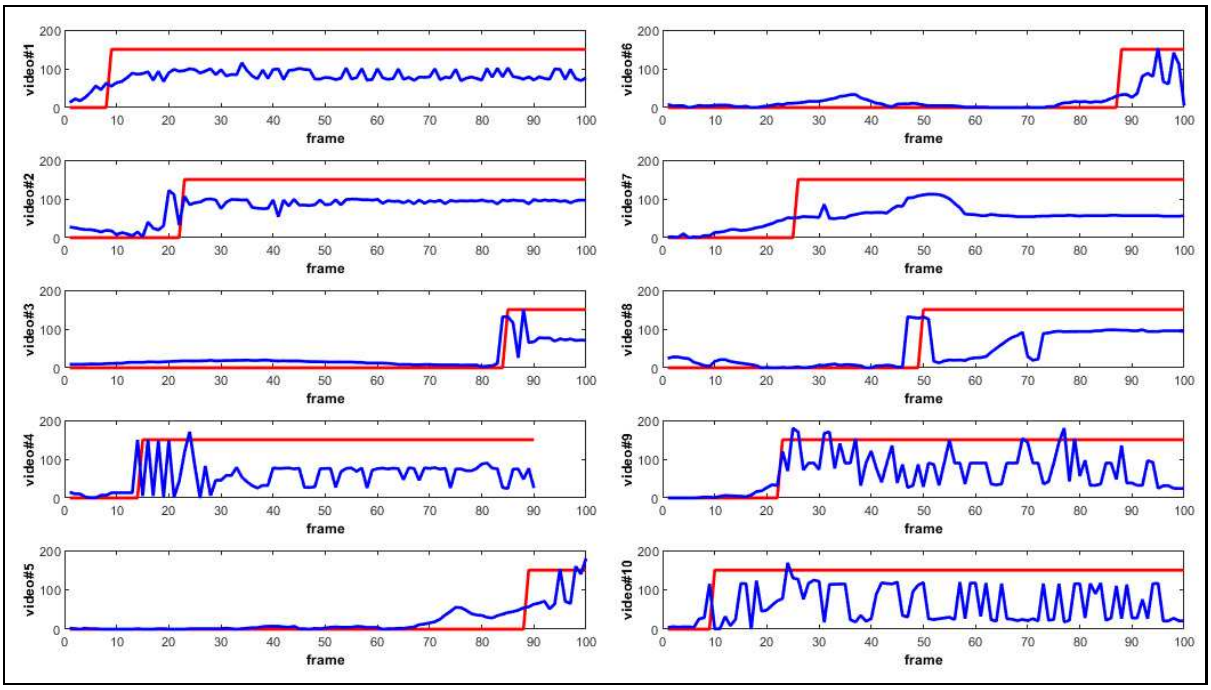

FIGURE 2. Samples of plotting the body orientation angle (the blue line) against the ground truth (the red line). It is obvious to see the difference between fall (squiggly line) and non-fall (straight-like line).

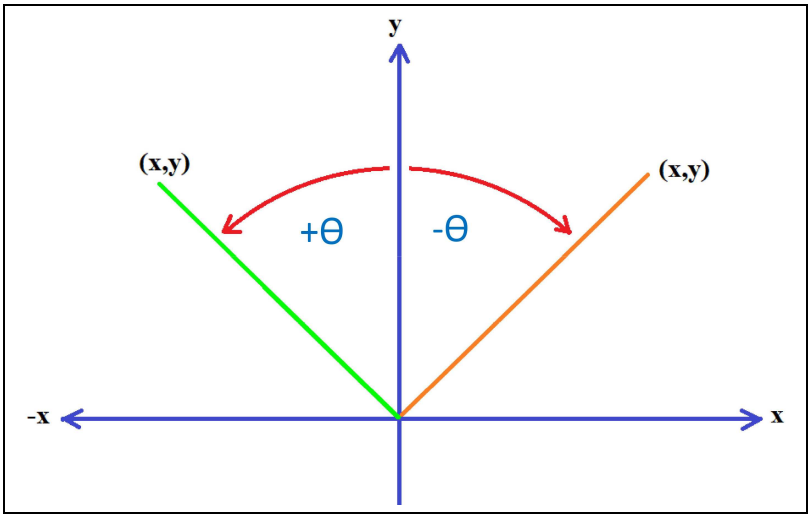

FIGURE 3. The angle between the Y-axis and the vector (x, y)

Figure 5. We subsequently draw a line on those key points. The calculation of the line against the Y-axis (θ) is shown in Figure 3, resulting in the body orientation angle, which we used as the input feature vector for SVM. The calculation of the body orientation angle is expressed in Equation (1).

Open-Pose delivers good results in extracting the human key points; however, the more key points we used to represent the upper human body, the less complete information is obtained. Therefore, we decided to pick only two key points representing the upper part of the body, to reduce the missed information while detecting. In addition, feature reduction can reduce the computational cost. Figure 2 demonstrates samples of plotting the body orientation angle (θ) (the blue line) against the ground truth (the red line). It is obvious to see that the body orientation angle (θ) difference between fall (squiggly line) and non-fall (straight-like line).

$$\theta_{rad} = \begin{cases} atan2(y,x) - \dfrac{\pi}{2} & \text{while } y > 0 \\ atan2(y,x) + \dfrac{3\pi}{2} & \text{while } y < 0 \text{ and } x < 0 \\ atan2(y,x) - \dfrac{\pi}{2} & \text{while } y < 0 \text{ and } x > 0 \end{cases} \tag{1}$$

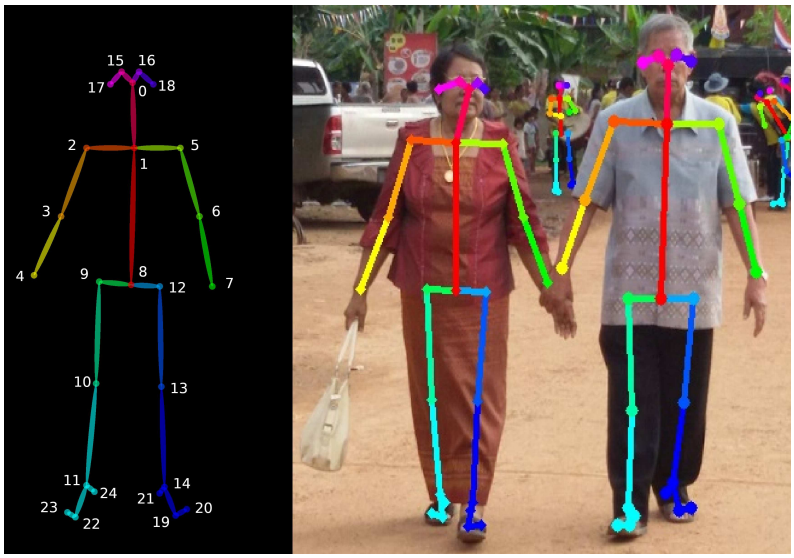

FIGURE 4. The left image is the OpenPose body 25 keypoints. The right image is the human pose skeletons result from OpenPose.

FIGURE 5. A sample result from Open-Pose skeleton extraction

2.2. **SVM classification.** Our fall detection method can be seen as a binary class classification problem on which a classifier has to decide if parts of individual frames represent a fall. Since we have an angle input, the well-known classifier SVM with Radial Basis Function (RBF) kernel can utilize them as an input feature. Before the training step, we have to find the optimum parameters, determined by grid search techniques, to avoid an "over-fitting" problem. We then apply the 10-fold cross-validation as shown in Figure 6, splitting the training set into ten subsets. For example, a model will be trained using 9 of folds as training data, and the result will be validated in the rest of the data repeating validation until ten loops. The performance measure reported by 10-fold cross-validation is the average of the values computed in the loop.

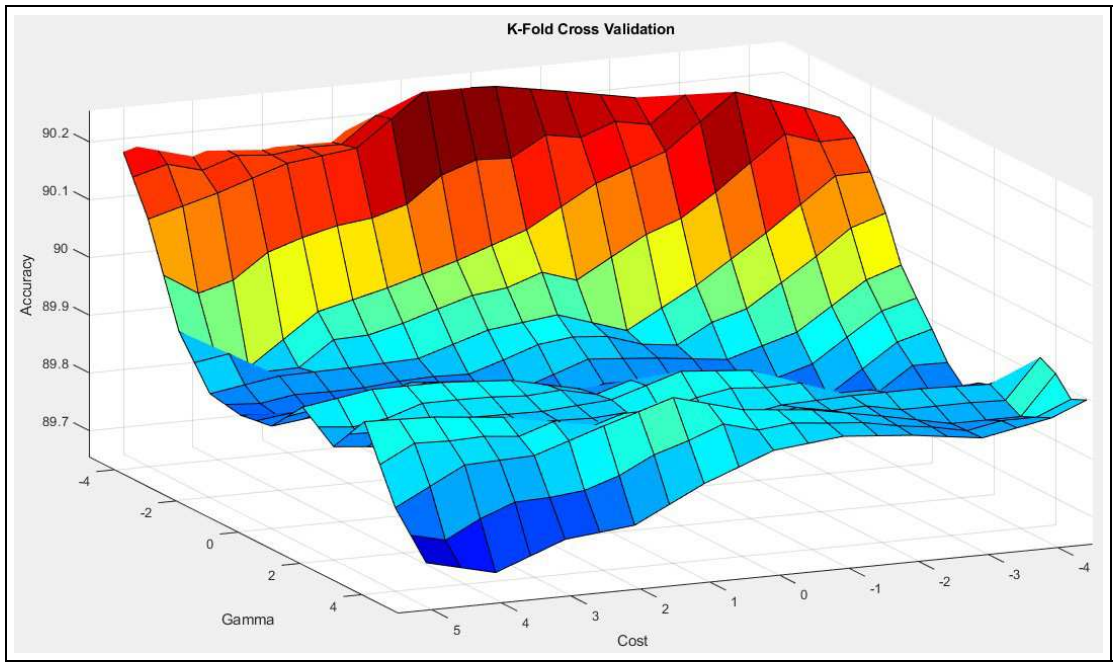

FIGURE 6. (color online) The plot is a heatmap of the classifier's cross-validation accuracy as a function of cost and gamma, which are 2 and 0.0625, to obtain the accuracy of 90.28%.

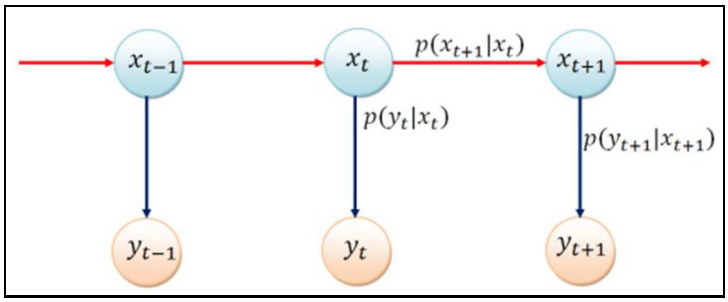

FIGURE 7. A graphical representation of a first order Hidden Markov Model. The state at time t, x_t, is dependent only on the state at the previous time step, x_{t-1}, and the current observation, y_t.

2.3. **Person detection and tracking.** To track persons in the video, we have first to locate them. In our work, the YOLOv4 was chosen because it delivered the best results for person detection [23]. YOLO is a single-CNN (Convolutional Neural Network) and can detect objects in real time. It provides the confidence value of the objects and draws a bounding box for each object. In our research, it is crucial to track the person because there is more than one person in a frame. The Open-Pose gives only the information of the current frame but does not provide the tracking information.

To track objects in video, we use a state-space model in which the person is modeled in each frame by a simple rectangle. The hidden state's transition from one observation time to the next and the observed video frame are related by a Hidden Markov Model [22], see Figure 7. We represent the hidden state at time t by $x_t = (c_x, c_y, \gamma, h, v_x, v_y, v_\gamma, v_h)$, where c_x and c_y are the bounding box center coordinate of the rectangle, γ is the aspect ratio of rectangle, h is the height of rectangle and v_x, v_y, v_γ, v_h are velocity variation value. The transition of the state from one time to another is represented by the state transition probability $p(x_{t+1}|x_t)$ and the probability of observing a frame y_t, given the state x_t is $p(y_t|x_t)$. As is well-known, tracking of the hidden state can be carried out using sequential Bayesian estimation (e.g., [10]), in which, given a probability density for the state having

made observations up to time t, $p(x_t|y_1, \ldots, y_t) \equiv p(x_t|Y_t)$, the prediction of the state at time $t+1$ is

$$p(x_{t+1}|Y_t) = \int p(x_{t+1}|x_t)p(x_t|Y_t)dx_t \tag{2}$$

Having observed a new video frame, the predicted state may be corrected using Bayes' rule:

$$p(x_{t+1}|y_{t+1}, Y_t) \propto p(y_{t+1}|x_{t+1})p(x_{t+1}|Y_t) \tag{3}$$

These prediction and correction steps may then be used sequentially to update the probabilities describing the hidden state to be updated as new data becomes available.

When both the state transition density $p(x_{t+1}|x_t)$ and the likelihood $p(y_t|x_t)$ are Gaussian, the prediction and correction steps become the well-known Kalman filter. In this work, we employ the deepSORT algorithm to track; deepSORT is an improved version of the Simple Online and Realtime Tracking (SORT) algorithm. This combination benefits from the tracking accuracy provided by the Kalman filter and the Hungarian algorithm and the appearance descriptor of the Deep Learning (CNN) Classifier. There are three components as follows. 1) Kalman Filter Estimator will predict the object's location based on the previous velocity. The result will be an Intersection over Union (IoU) Distance matric between the previous frame tracked and the Yolo detected bounding box. 2) A Deep appearance descriptor is a single visual feature vector extracted from CNN; Mahalanobis distance is used to determine the relationship of the detected object with the existing track. 3) Data association, Hungarian algorithm assign detected objects that match existing track using location and appearance metrics.

3. **Performances.** In this section, we describe the dataset we use in our work. Next, we give details of the experiment and the results.

3.1. **Datasets.** There are 346 colour videos from five different datasets used for our work, 214 videos for training, and 132 videos for testing.

1) FallAllD: A Comprehensive Dataset of Human Falls and Activities of Daily Living [26].

2) UP-Fall Detection Dataset (UPFD): There are 11 activities and three trials per activity. Subjects performed six simple human daily activities and five different types of human fall. These data were collected from over 17 healthy young adults without impairment using a multimodal approach, i.e., wearable, ambient, and vision devices [17].

3) UR-Fall Detection Dataset (URFD): The dataset contains 70 video sequences (30 falls + 40 activities of daily living). Fall events are recorded with 2 Microsoft Kinect cameras and corresponding accelerometric data. ADL events are recorded with only one device (camera 0) and an accelerometer. Sensor data was collected using PS Move (60Hz) and x-IMU (256Hz) devices [13].

4) ImViA Fall Detection dataset (ImViA): The dataset contains 191 videos we annotated. The videos are recorded from different locations, allowing for the definition of several evaluation protocols (Home, Coffee room, Office, and Lecture room). The frame rate is 25 frames/s, and the resolution is 320×240 pixels various normal daily activities and falls [14].

5) Multiple Cameras Fall Dataset (MCFD): This dataset contains 24 scenarios recorded with 8 IP video cameras. The first 22 scenarios contain a fall and confounding events; the last two ones contain only confounding events [6].

3.2. **Result and discussion.** Our experiment is performed on five datasets and contains 214 videos in realistic recorded from a video surveillance camera. The frame rate is 25 frames/s, and the resolution is 320×240 pixels. Every single frame, there is a person who may fall or may not. So, we annotated in which a person was detected as either falling or not falling. Figures 8 and 9 illustrate some of the results.

FIGURE 8. Illustration of some images of our experiment results in the coffee room environment

To analyze the detection results of the proposed method, the following parameters are used.

- True Positive (TP) is the number of fall events detected correctly.
- True Negative (TN) is the number of non-fall events detected correctly.
- False Positive (FP) is the number of non-fall events detected as fall events.
- False Negative (FN) is the number of fall events detected as non-fall events.

The correct classification rate, the ratio of correct case (TP and TN) in the population, is expressed in Equation (4)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

The sensitivity, the capacity to detect fall events, is expressed in Equation (5)

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

The specificity, the capacity to detect non-fall events, is expressed in Equation (6)

$$Precision = \frac{TN}{TN + FP} \tag{6}$$

The error rate, the incorrect classification rate, is expressed in Equation (7)

$$Error = \frac{FP + FN}{TP + FP + TN + FN} \tag{7}$$

FIGURE 9. Some images of our experiment results in the coffee home environment

Table 1 presents the results of our work. There are five datasets: Dataset 1 (FallAllD), Dataset 2 (UPFD), Dataset 3 (URFD), Dataset 4 (ImViA) and Dataset 5 (MCFD). We obtain 89.66% average classification accuracy and 90.78% and 88.27% sensitivity and specificity, respectively. The average accuracy is calculated by summing the number of correctly classified values (true positive) and dividing it by the total number of values, resulting in 89.66%.

TABLE 1. The results of fall detection

Dataset	Accuracy (%)	Recall (%)	Precision (%)	Error (%)
FallAllD	92.82	99.92	83.04	7.17
UPFD	89.65	100	71.23	10.34
URFD	93.77	99.07	92.10	6.22
ImViA	90.28	83.42	94.29	9.71
MCFD	87.57	91.17	77.07	12.42
Average	89.66	90.78	88.27	10.33

4. **Conclusions.** In conclusion, this paper proposed a vision-based fall detection combining human-pose and SVM as the method of feature augmentation. Firstly, we used the Open-Pose to extract the 25 human key points; we then picked two essential key points: neck and mid-abdomen, to represent the upper human body. Next, we calculate the body orientation angle against the Y-axis used as the SVM input feature vector. In the testing

step, we work parallel YOLOv4 with deepSORT and Open-Pose to track multiple persons. We obtain 89.66% average classification accuracy. In the future, we may conduct more experiments on fall detection in various environments, for example, at night or in the outdoor environment, employing the 3D human pose to achieve higher accuracy.

Acknowledgment. This research was supported by Faculty of Science and Engineering, Kasetsart University, Chalerm Phra Kiat Sakon Nakhon Province Campus, Thailand. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] T. E. Kennedy, The prevention of falls in later life. A report of the kellogg international work group on the prevention of falls by the elderly, *Danish Medical Bulletin*, vol.34, no.4, pp.1-24, 1987.

[2] The Cambridge Online Dictionary, *Fall*, 2021, https://dictionary.cambridge.org/dictionary/english/fall?q=Fall, Accessed on May 22, 2021.

[3] K. Adhikari, H. Bouchachia and H. Nait-Charif, Deep learning based fall detection using simplified human posture, *International Journal of Computer and Systems Engineering*, vol.13, no.5, pp.251-256, 2019.

[4] M. Alonso, A. Brunete, M. Hernando and E. Gambao, Background-subtraction algorithm optimization for home camera-based night-vision fall detectors, *IEEE Access*, vol.7, pp.152399-152411, 2019.

[5] M. Alwan, P. J. Rajendran, S. Kell, D. Mack, S. Dalal, M. Wolfe and R. Felder, A smart and passive floor-vibration based fall detector for elderly, *2006 2nd International Conference on Information Communication Technologies*, vol.1, pp.1003-1007, 2006.

[6] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud and J. Rousseau, *Multiple Cameras Fall Dataset*, Technical Report, 2010.

[7] Z. Bian, J. Hou, L. Chau and N. Magnenat-Thalmann, Fall detection based on body part tracking using a depth camera, *IEEE Journal of Biomedical and Health Informatics*, vol.19, no.2, pp.430-439, 2015.

[8] R. Cucchiara, A. Prati and R. Vezzani, A multi-camera vision system for fall detection and alarm generation, *Expert Systems*, vol.24, no.5, pp.334-345, 2007.

[9] A. Doulamis and N. Doulamis, Adaptive deep learning for a vision-based fall detection, *Proc. of the 11th PErvasive Technologies Related to Assistive Environments Conference (PETRA'18)*, New York, NY, USA, pp.558-565, 2018.

[10] M. Ghosh, N. Mukhopadhyay and P. K. Sen, Sequential Bayesian estimation, *Sequential Estimation*, pp.111-152, 1997.

[11] R. J. Gurley, N. Lum, M. Sande, B. Lo and M. H. Katz, Persons found in their homes helpless or dead, *New England Journal of Medicine*, vol.334, no.26, pp.1710-1716, 1996.

[12] J. He, M. Zhou, X. Wang and Y. Han, Application of kalman filter and k-NN classifier in wearable fall detection device, *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp.1-7, 2017.

[13] B. Kwolek and M. Kepski, Human fall detection on embedded platform using depth maps and wireless accelerometer, *Computer Methods and Programs in Biomedicine*, vol.117, pp.489-501, 2014.

[14] ImViA Laboratory, *Fall Detection Dataset*, 2020, https://imvia.u-bourgogne.fr/en/database/fall-detection-dataset-2.html, Accessed on May 22, 2021.

[15] C.-L. Liu, C.-H. Lee and P.-M. Lin, A fall detection system using k-nearest neighbor classifier, *Expert Systems with Applications*, vol.37, no.10, pp.7174-7181, 2010.

[16] N. Lu, Y. Wu, L. Feng and J. Song, Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data, *IEEE Journal of Biomedical and Health Informatics*, vol.23, no.1, pp.314-323, 2019.

[17] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez and C. Peñafort-Asturiano, Up-fall detection dataset: A multimodal approach, *Sensors*, vol.19, no.9, DOI: 10.3390/s19091988, 2019.

[18] M. Yu, S. M. Naqvi and J. Chambers, Fall detection in the elderly by head tracking, *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pp.357-360, 2009.

[19] A. Núñez-Marcos, G. Azkune and I. Arganda-Carreras, Vision-based fall detection with convolutional neural networks, *Wireless Communications and Mobile Computing*, 2017.

[20] World Health Organization, *Falls*, 2021, https://www.who.int/news-room/fact-sheets/detail/falls, Accessed on May 22, 2021.

[21] Y. Raaj, H. Idrees, G. Hidalgo and Y. Sheikh, Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4620-4628, 2019.

[22] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. of the IEEE*, vol.77, no.2, pp.257-286, 1989.

[23] J. Redmon, S. K. Divvala, R. B. Girshick and A. Farhadi, You Only Look Once: Unified, real-time object detection, *CoRR*, abs/1506.02640, 2015.

[24] C. Rougier, J. Meunier, A. St-Arnaud and J. Rousseau, Robust video surveillance for fall detection based on human shape deformation, *IEEE Trans. Circuits and Systems for Video Technology*, vol.21, no.5, pp.611-622, 2011.

[25] S. Sadigh, A. Reimers, R. Andersson and L. Laflamme, Falls and fall-related injuries among the elderly: A survey of residential-care facilities in a Swedish Municipality, *Journal of Community Health*, vol.29, pp.129-140, 2004.

[26] M. Saleh and R. Le Bouquin Jeannes, FallAllD: A comprehensive dataset of human falls and activities of daily living, *IEEE Dataport*, DOI: 10.21227/bnya-mn34, 2020.

[27] Thailand Public Health Statistics, *Public Health Statistics*, 2018, https://bps.moph.go.th/new_bps/sites/default/files/statistic%2061.pdf, Accessed on April 22, 2021.

[28] V. Vaidehi, K. Ganapathy, K. Mohan, A. Aldrin and K. Nirmal, Video based automatic fall detection in indoor environment, *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, pp.1016-1020, 2011.

[29] Y. Wang, W. Xiong, J. Yang and S. Wang, A new fall detection method based on fuzzy reasoning for an omni-directional walking training robot, *International Journal of Innovative Computing, Information and Control*, vol.16, no.2, pp.597-608, 2020.

[30] D. Wild, U. S. Nayak and B. Isaacs, How dangerous are falls in old people at home?, *Br. Med. J. (Clin. Res. Ed.)*, DOI: 10.1136/bmj.282.6260.266, 1981.

[31] A. Williams, D. Ganesan and A. Hanson, Aging in place: Fall detection and localization in a distributed smart camera network, *Proc. of the 15th ACM International Conference on Multimedia*, pp.892-901, 2007.