Neural Computing and Applications (2023) 35:6515–6528
https://doi.org/10.1007/s00521-022-08028-0

ORIGINAL ARTICLE

Identifying human activities in megastores through postural data to monitor shoplifting events

Mohd. Aquib Ansari[1] · Dushyant Kumar Singh[1]

Received: 20 April 2022 / Accepted: 29 October 2022 / Published online: 22 November 2022

Abstract

In recent years, modeling activity patterns for understanding events and human behavior has drawn prominent attention in research. Multiple methods have been proposed for developing automated vision systems that are capable of inferring accurate semantics from the moving dynamics. The multi-disciplinary nature of Human Activity Recognition (HAR) methods and the expanding technologies in this field inspire continual updates in existing methods. However, a cost-effective solution is still needed to recognize human activities like shoplifting in an occluded environment. With this motivation, we present a novel approach to identify human stealing actions by analyzing the postural information of the human body. This approach involves extracting 2D postural body joints of a human being from the captured frame. Pose encoding and postural feature generation in parameter space are the foremost contributions of this work, which can handle the occluded actions too. The feature reduction is done to scale the features into a smaller dimension with an objective of the computationally efficient and real-time solution. Activity classification is done on the reduced feature sets to detect human shoplifting actions in real-time scenarios. Experiments are performed on the synthesized shoplifting dataset, where the results derived are found more promising compared to other state-of-the-art methods, with an accuracy of 96.87%. Additionally, this method exhibits commendable real-time performance in processing actual store camera footage.

Keywords Surveillance system · Posture-based action recognition · Human pose encoding scheme · Spatiotemporal features · Deep neural network

1 Introduction

Nowadays, shoplifting is a major concern of retail outlets. It is an act of secretly picking and hiding goods from an open retail establishment with the intention of not paying for them. Traders suffer huge losses due to shoplifting and those who are involved in such acts are called shoplifters. According to the report [1], many supermarkets face monthly losses between Rs. 50,000 and Rs. 1 lakh due to shoplifting. Here, shopkeepers reported that women outnumber men in shoplifting and some even incite their children to commit crimes. Despite the widespread adoption of CCTV-based video surveillance infrastructure by the storekeepers, it is still challenging to trace shoplifters while committing shoplifting. In the existing infrastructure, the control room personnel attentively examine the video stream received from CCTV infrastructure. Examining massive video footage for hours leads to fatigue and loss of attention, causing security personnel might miss to looking at the critical frames where shoplifting took place. As a result, video surveillance systems in use today are classified as passive or non-intelligent video surveillance. The surveillance that is done passively defeats the objective of reducing crime in stores/shops. Therefore, an automated and proactive video surveillance system is in serious demand to seamlessly examine video feeds and produce an alert while anyone is involved in the act of shoplifting. The generated alert from such a system on detecting the shoplifting activity can help the security personnel to take retrofitting action in time.

✉ Mohd. Aquib Ansari
mansari@mnnit.ac.in

Dushyant Kumar Singh
dushyant@mnnit.ac.in

[1] CSED, MNNIT Allahabad, Prayagraj, Uttar Pradesh, India

In order to get this objective, only few of the researchers have worked on it. First, Arroyo et al. [2], in 2015, proposed a passive framework to prevent shoplifting by detecting loitering events. In 2021, Guillermo et al. [3] identify human stealing actions using a three-directional convolutional neural architecture that encodes spatial and temporal characteristics for an action. In 2022, Ansari and Singh [4] learn spatiotemporal dynamics for stealing actions representation by taking the benefit of both convolutional neural network (CNN) and recurrent learning classifier. However, chaotic background, occluded postures and variation in human pose create difficulty in detecting shoplifting, causing loss in detection accuracy. Therefore, a cost-effective solution is still demanded to deal with these scenarios and catch the shoplifters when they commit theft.

Human activity recognition (HAR) [5, 6] has attracted a lot of attention in recent years because many applications have their role either directly or indirectly. These applications include security and surveillance, healthcare monitoring, human–computer interaction, assisted living, smart homes, and more. The methods [7, 8] used in the existing HAR systems typically adopt handcrafted and deep-learning features that are determined from RGB and optical flow encoded video sequences. These methods have ability in identifying/classifying the activities but may fail in differentiating activities in case of body pose variations and occluded environment.

Similarly, human posture estimation methods [9–11] have advanced significantly with the development of more robust algorithms. Human body posture represents high-level semantics with discriminative geometric connections between articulated body joints. Therefore, human posture estimation-based algorithms have widely popular in video-processing-based recognition interfaces. Activity recognition is highly dependent on the accurate estimation of human pose. Traditional pose estimation approaches were based on graph modeling and optimization, which take a longer time in posture generation. Recently, the OpenPose [12] algorithm has been designed to accurately detect body joints with excellent computational performance. It detects key points for facial appearance, body and arm joints by inputting RGB images through advanced deep neural architecture.

Although HAR systems are customarily deployed in most video classification-based tasks, there is still a lot of research to be needed to improve accuracy, reduce computation costs, and deal with occluded environments. To allay these apprehensions and create a cost-effective solution, we focus on developing an advanced 2D pose-based activity recognition method using deep neural architecture to access complex activities like shoplifting in real-time scenarios. Here, 2D skeleton-based human postures are analyzed to identify human stealing acts in real-time scenarios. The geographical location of body joints is estimated using the OpenPose algorithm. Pose encoding in the parameter space and extracting relevant features to represent a human act uniquely are the main contribution of this work. The extracted features are scaled into lower dimensional space using principle component analysis (PCA) and passed to fully connected classification layers of deep neural network (DNN) to categorize the video sequences into respective classes.

The following is a list of major contributions:

- We develop an activity recognition framework that jointly estimates 2D poses and distinguishes associated actions over synthesized training inputs.
- We present customary pose encoding and feature evaluation schemes to extract pertinent features from postural/skeleton data sets.
- We perform a wide range of experiments on a shoplifting dataset developed by CV Laboratory of MNNIT Allahabad.
- Additionally, we assess the performance of the proposed method for various real-store camera footage.

The remainder of the article is arranged as follows. Section 2 includes the interrelated research of existing HAR systems. The proposed methodology with pose encoding and feature evaluation schemes is discussed in Sect. 3. The experimental results are part of Sect. 4, followed by conclusions in Sect. 5.

2 Literature survey

Activity recognition has been comprehensively investigated during the past few decades. Understanding human behavior is critical to the high-level interaction between computers and humans by embodying more intelligent systems. Most of the intelligence systems use convolutional neural networks (CNN) to analyze human behavior in real-time surveillance. Learning temporal dynamics is an arduous task and prior techniques have incorporated handcrafted features like human silhouette, Histogram of Oriented Optical Flow (HOF), Histogram of Oriented Gradient (HOG) etc., to design a wide range of information-encoded descriptors. Sanal Kumar and Bhavani [13] proposed an activity recognition framework using HOG, GiST and color features in multimodal egocentric videos. This work uses a random forest classifier to categorize the extracted feature sets into their respective classes. Guillermo et al. [3] use an advanced CNN network named three-dimensional convolutional neural network architecture (3D-CNN) for activity classification tasks. They found that the 3D-CNN features can encode contextual information more accurately than the 2D-CNN.

Recurrent neural networks (RNN) [4, 14] have lately been demonstrated to perform well in machine translation, voice recognition, activity recognition, image and video description. Long Short-Term Memory (LSTMs) and Gated Recurrent Units (GRUs) are the advanced archetypes of RNN networks to deal with sequence prediction problems more accurately. The computing power of GRUs is much faster than LSTMs. However, LSTMs can learn longer sequences than GRUs and play a great role in human activity classification tasks. Most existing activity detection methods extract features using CNN, perform feature pooling and classify sequences using the LSTM network. Ansari and Singh [4] take benefit of CNN and the recurrent learning process to classify human acts in indoor surveillance. Deep Inception V3 architecture is utilized here to mine pertinent features from the spatial stream and the LSTM network is used for performing classification tasks. Some existing HAR systems use kinetics together with spatial information to accurately classify human actions. The extended research [7] proposed various proposals to compose motion dynamics from RGB sequences and provided their consequence outcomes to access shoplifting scenarios for different modeling parameters. The motion stream is composed here by combining the optical flow vectors and the HOG vectors for each sequence. Deep Inception V3 is used to extract relevant features from spatial and motion streams and classification is done using LSTM network. Jayaswal and Dixit [15] proposed an anomalies classification framework for synthesized training inputs. The framework uses a modified Xception model to extract features from video sequences and LSTM network for anomalies classification. Donahue et al. [16] proposed a large-scale visual understanding framework using recurrent convolutional networks. The framework uses the recurrent sequence model associated with the modern convolution model that learns the spatiotemporal dynamics for the visuals more accurately.

The methods as discussed above being image structure-based methods to recent CNN approaches, use visual semantics to propose HAR systems. Additionally, some HAR systems are now adopting posture estimation methods to understand human behavior from video sequences. Posture estimation addresses the problem of locating the structural key points of body components. The techniques used in pose estimation can solve many issues in several areas, such as augmented reality, action recognition and animation gaming. Recently, most of the posture estimation methods use CNN architecture to identify the body's joints in real-time. Bottom-up and top-down are the two approaches used in 2D pose estimation. The bottom-up approach first traces the body joints and then assorts them to create unique postures, whereas the second approach first traces the individuals and then locates the joints for each detected individual.

Noori et al. [17] proposed an activity recognition system using 2D pose estimation in a deep recurrent process. The system uses the OpenPose algorithm to extract the 2D pose of humans, motion features to develop the motion dynamics. The long short-term memory (LSTM) network is then used to analyze actions having previously extracted features as the prominent input. Wang and Wu [18] proposed an expert activity recognition framework for complex action recognition by conjoining three deep designs of action motions, objects and scene semantics representation. These deep designs are used to capture both contextual and temporal information of actions. Here, an lp-norm multiple kernel learning method combines these three deep cues and learns classifiers of actions. Li et al. [11] use deep neural architecture to model human 2D posture for action recognition. The framework encodes body poses into multiple representations and computes CNN features from each of them. The computed features from each presentation are aggregated into two vectors and Softmax classification is used to classify each video. The final score is evaluated over pose and global score through the weighting layer to label the activity present in extracted features.

As seen in the literature, several researchers have contributed their research to the area of human behavior analysis. However, a cost-effective solution is still required to distinguish complex acts in occluded environments. Above it, the area of identifying shoplifting activities in occluded environment is still found to be untouched. Therefore, this paper proposes a framework for recognizing shoplifting using human postural data analysis in normal and occluded environments.

3 Proposed methodology

This work focuses on identifying acts of shoplifting by using human posture related data. Figure 1 presents the block diagram to illustrate the overview of the proposed methodology. The method uses video sequences/frames captured from the CCTV camera as input and passes them to the posture generation module. The posture generation module extracts skeletal information from the input sequence along with the spatial location of the body joints for the individual. The extracted skeleton information containing body joints is made input to the proposed pose encoding scheme, where four important tasks, i.e., discarding irrelevant frames, choosing selective joints, coordinate scaling and estimation of missing joints are performed. In the next step, relevant features like normalized joints, motion features, angle and distance are calculated from the N encoded skeletons, which are

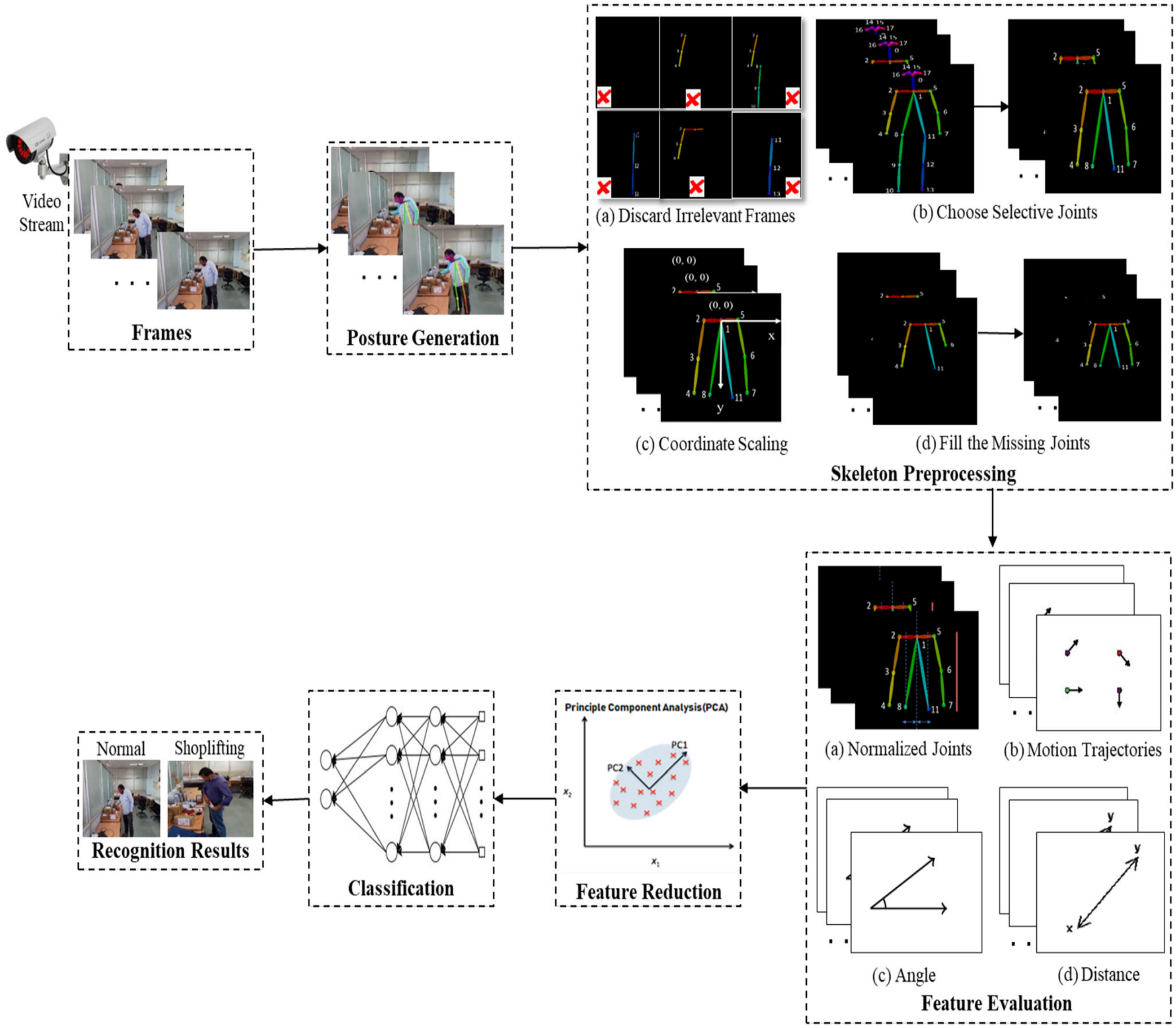
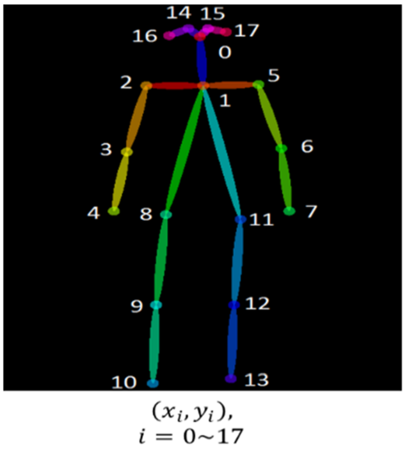


Fig. 1 Block diagram of the proposed methodology

subsequently used by the feature reduction process. The feature reduction process transforms the large-size feature set into compact-sized features by suppressing redundant and insignificant information. Finally, the transformed features are passed to feed-forward deep neural model to classify these sequences into respective classes (i.e., normal and shoplifting). The discussion on the above-mentioned technicalities is provided in the upcoming subsections.

3.1 Posture generation

Posture generation is an essential module of this work that uses OpenPose algorithm to extract skeleton information for a person. OpenPose localizes the 2D body joints present in the human skeleton. In this algorithm, first feature maps are extracted from inputted RGB images through a baseline CNN. A multi-stage CNN then evaluates Part Confidence Maps (PCM) and Part Affinity Field (PAF). The belief behind PCM is that any given pixel can specify a particular body component in two-dimensional space. Other hand, PAF comprises a series of 2D vectors that encode the orientation and position of people's limbs in the image. The last stage parses the confidence maps and affinity fields using a greedy bipartite matching algorithm to get the 2D key points of each person in an image. The proposed solution extracts the spatial location of 18 body joints using OpenPose, as shown in Fig. 2. Further, the extracted joint's positions are analyzed to identify the action's performed by an individual in video sequences.



0 :	Nose	9 :	Right Knee
1 :	Neck	10 :	Right Ankle
2 :	Right Shoulder	11 :	Left Hip
3 :	Right Elbow	12 :	Left Knee
4 :	Right Wrist	13 :	Left Ankle
5 :	Left Shoulder	14 :	Right Eye
6 :	Left Elbow	15 :	Left Eye
7 :	Left Wrist	16 :	Right Ear
8 :	Right Hip	17 :	Left Ear

Fig. 2 OpenPose skeleton joints

3.2 Skeleton pre-processing or pose encoding

The raw skeleton data consisting of 18 body joints is not directly used for prominent feature extraction to avoid inaccurate training and inconsistent outcomes. A four-step pose encoding is therefore proposed here to polish the prominent feature extraction process. Here, pose encoding is done in four steps which are discussed as follows:

a. **Discard irrelevant frames:** It first discards those frames where OpenPose detects no human skeleton or the detected skeleton has no neck or thigh.
b. **Remove all head joints and leg joints:** The human skeleton generated by OpenPose represents 18 body joints. However, the position of the head and legs contributes very less to the shoplifting kind of activities and so we remove five points of the head (i.e., one head, two ears and two eyes) and two points of the legs (i.e., two ankles and two knees) from the skeleton data. This will help in reducing feature size and overall computability. The final skeleton has nine body joints: one neck, two shoulders, two elbows, two wrists, and two hips, as shown in Fig. 3.

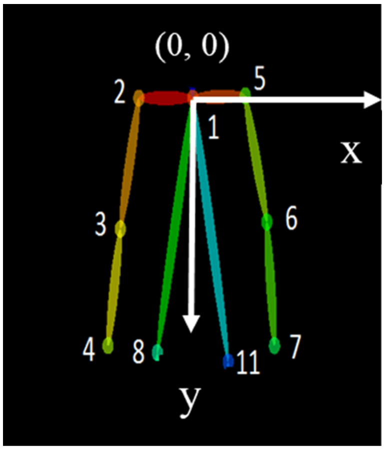


Fig. 4 Scaling coordinates

c. **Coordinate Scaling:** The joint’s position obtained by the preceding step comprises different units of x and y coordinates, therefore, scaling is required to transform these coordinates into the same units. Algorithm 1 shows the coordinate scaling procedure, where it takes a skeleton (S) with sets of body joints as input and produces scaled body joints as output. The scaling positions the neck joint at the origin (0, 0) by subtracting the coordinate value of the neck joint from

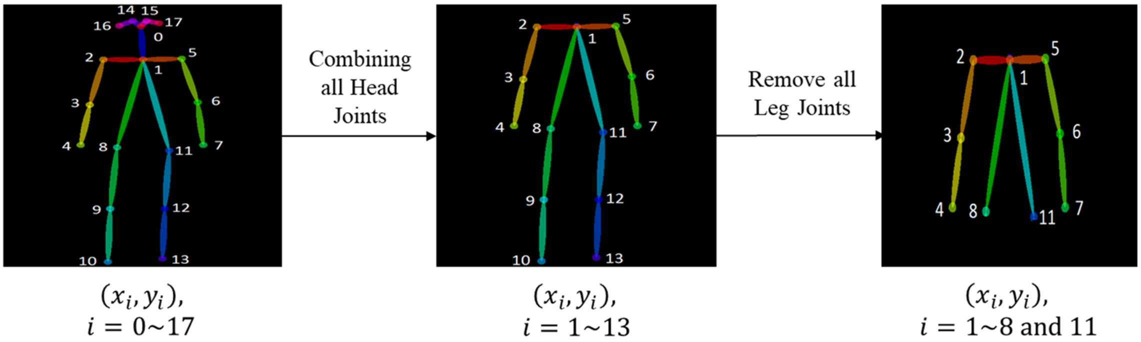


Fig. 3 Joint configuration

the coordinate values of all the body joints. Figure 4 shows scaled body joints at the 2D plane.

Algorithm #1: Coordinate Scaling

Input: Skeleton S with selected i body joints

$\mathbf{S = \{x_0, y_0, x_1, y_1, x_2, y_2 \quad \ldots, x_i, y_i\}}$

Output: Scaled coordinate of skeleton S

```
Function Scale_body_offset(S):
    S = S.copy()
    px0, py0 = get_joint(S, NECK)
    S[0 : : 2] = S[0 : : 2] - px0
    S[1 : : 2] = S[1 : : 2] - py0
    Return S
```

d. **Fill the missing joints:** Sometimes, OpenPose fails to discover a widespread skeleton from the image due to the occlusion or camera viewpoint, which causes some blank space on the missing body joints. In order to preserve a fixed-size feature vector, the missing joints in current frame are calculated through their relative positioning w.r.t. the previous frames in the sequence. Equations (1) and (2) present the calculation of missing joint coordinates (x_i, y_i).

$$x_i = x_0 + x_i\_\text{prev} \tag{1}$$

$$y_i = y_0 + y_i\_\text{prev} \tag{2}$$

Here (x_0, y_0) is the spatial location of the neck joint in the current frame and $(x_i\_\text{prev}, y_i\_\text{prev})$ is the set of the spatial locations of the missing joints in the previous frame.

3.3 Features evaluation

This section presents the proposed feature-encoding schemes using pose encoded skeleton data containing nine joints each. Here, a window of size N is taken to represent N preprocessed skeleton frames that encode a complete human act. Let $X = \{S_0, S_1, S_2, S_3, \ldots, S_N\}$, where X is the set of skeletons for N frames. Each point in the skeleton is represented by pair of spatial coordinates (x_i, y_i). To get unique features from the skeletons and their body parts, the feature's calculation using N frames is explained as given below:

a. **Normalized body joints:** The normalized body joints $(Xnorm)$ are calculated by normalizing each body joint for N skeletons. It is so done by dividing the scaled body offset by the mean height of N skeletons, as shown in Eq. (3). The height of each skeleton is calculated as explained in **Algorithm 2**, where the distance between the neck and hips is calculated using Euclidean distance.

$$Xnorm = \frac{S_j}{\text{Mean(Height H of each skeleton)}}, \tag{3}$$

Where $0 \leq j \leq N$

Algorithm #2: Get Skeleton Height H

Input: Skeleton S with selected i body joints

$\boldsymbol{S = \{x_0, y_0, x_1, y_1, x_2, y_2 \quad \ldots, x_i, y_i\}}$

Output: Height H of the skeleton

```
Function Get_Body_Height(S):
    $x0, y0 \leftarrow Get\_Joint(S, NECK)$
    $x11, y11 \leftarrow Get\_Joint(S, L\_HIP)$
    $x12, y12 \leftarrow Get\_Joint(S, R\_HIP)$
    $if\ y11 == NaN\ and\ y12 == NaN:$
        $\boldsymbol{Return}$ 1.0
    $if\ y11 == NaN:$
        $x1, y1 \leftarrow x12, y12$
    $elif\ y12 == NaN:$
        $x1, y1 \leftarrow x11, y11$
    $else:$
        $x1, y1 \leftarrow \frac{(x11 + x12)}{2}, \frac{(y11 + y12)}{2}$
    $H \leftarrow \sqrt[2]{(x0 - x1)^2 + (y0 - y1)^2}$
    $\boldsymbol{Return}\ H$
```

b. **Motion between body joints:** The motion between body joints (V) is calculated by subtracting the joint position in the current frame from the joint position in the previous frame, as shown in Eq. (4).

$$V = X[S_N] - X[S_{N-1}] \tag{4}$$

c. **Angle between body joints:** Assume (x_j, y_j) and (x_k, y_k) are the two points in the skeleton, the angle (θ) between two body points can be obtained through Eq. (5). The body points for which angles are calculated are presented in Table 1.

Table 1 Angle and distance calculation between points

Angle (θ) and distance (L)	Part 1	Part 2
1	Neck	Right shoulder
2	Neck	Left shoulder
3	Right shoulder	Right elbow
4	Left shoulder	Left elbow
5	Right elbow	Right wrist
6	Left elbow	Left wrist
7	Right wrist	Right hip
8	Left wrist	Left hip

$$\theta = \tan^{-1}\frac{\left(x_j - x_k\right)}{\left(y_j - y_k\right)} \tag{5}$$

d. **Length of the limb:** The length of the limb (L) is simply the Euclidian distance between two body points and is calculated by Eq. (6). Table 1 presents the body points for which length parameter between two body joints, i.e., $\left(x_j, y_j\right)$ and $\left(x_k, y_k\right)$, or limb is calculated.

$$L = \sqrt[2]{\left(x_k - x_j\right)^2 + \left(y_k - y_j\right)^2} \tag{6}$$

Finally, extracted features obtained from the above-mentioned feature encoding schemes are concatenated to represent final features, as shown in Eq. (7). Further, these concatenated features are processed for dimensionality reduction.

$$\text{Features} = [Xnorm,\ V,\ \theta,\ L] \tag{7}$$

The summary related to extracted features is presented in Table 2. As the value of N is assigned to be 145, the dimensions of extracted features obtained by $Xnorm$, V, θ and L are 2610, 2592, 1160 and 1160, respectively. Therefore, the total feature length to represent an action in N frames is 7552. Further, this feature set is supplied to the dimensionality reduction module to select prominent features.

3.4 Dimensionality reduction

Dimension reduction is a data formation/refinement procedure done on the raw data prior to modeling. It eliminates the least significant features from the data as these might degrade the model's accuracy. This work uses Principal Component Analysis (PCA), an unsupervised feature transformation technique, for dimensionality reduction. PCA helps to overcome the overfitting problem by reducing the number of features. It is common to speed up machine learning algorithms by eliminating correlated variables that do not contribute to any decision-making. PCA identifies patterns in data based on the correlation between features and transforms them into new subspace with a lower dimensionality without sacrificing any important information. In essence, it constructs an $m \times n$ dimensional transformation matrix W that maps a sample vector/features x onto a new k-dimensional feature subspace, as shown in Eq. (8).

$$\left.\begin{array}{ll} x = [x_1, x_2, \ldots, x_m], & x \in R^m \\ \quad \downarrow xW,\, W \in R^{(m \times n)} & \\ z = [z_1, z_2, \ldots, z_m], & z \in R^n \end{array}\right\} \tag{8}$$

As a result of remodeling m-dimensional input data into n-dimensional subspace (where $m \le n$), the principal components are arranged in order of descending variance values given the constraint that each principal component is uncorrelated to other components. Finally, these components are used to represent the features that are further used to build the deep neural architecture for classification.

3.5 Activity classification

The transformed features obtained from the dimensionality reduction process are more relevant and are now used for activity classification to categorize activities (i.e., shoplifting and normal) present in the video sequences. Following the deep CNN architecture, this paper replaces the functionality of convolution and max-pooling operations with our proposed scheme to extract pertinent features and then uses a fully connected feed-forward neural network (FFNN) architecture to learn and perform the classification task. FFNN is a supervised learning approach that learns a function $f(\cdot) : R^n \rightarrow R^o$ by training on a dataset. Here, n and o represent dimensions for input and output, respectively. Basically, the FFNN network consists of three main layers: an input layer, hidden layer and an output layer. The input layer consists of a set of neurons $\{f_i | f_1, f_2, \ldots, f_m\}$ representing input features. The hidden layer is the backbone of FFNN that performs all the computation tasks. There can be one or more than one hidden layer in FFNN network. Each neuron in the hidden layer transforms the vectors obtained from the input layer with a weighted linear summation $w_1 f_1 + w_2 f_2 + \cdots w_m f_m$, followed by a nonlinear activation function. Finally, the output layer turns the vectors obtained from the last hidden layer into output values.

Table 2 Features summary

Features	Representation	Evaluation	Vectors
Normalized body joints	$(Xnorm)$	9 Joints × 2 position/joint × N frames	2610
Motion between body joints	(V)	9 Joints × 2 position/joint × $(N - 1)$frames	2592
Angle between body joints	(θ)	8 Length × N frames	1160
Distance between body joints	(L)	8 Length × N frames	1160
Total feature vectors $[Xnorm, V, \theta, L]$			7522

The proposed solution utilizes FFNN with one input layer, three hidden layers and one output layer. The size of the input layer is 4800 that represents input features. The hidden layers I, II and III consist of 1024, 512 and 256 neurons, respectively. Other side, the output layer contains two neurons that represent two classes, namely normal and shoplifting. The FFNN is hyper tuned with Tanh activation function in hidden layers and sigmoid activation function in the output layer. Adam optimizer is used to optimize weights of the neural network and the learning rate is 0.001. The configuration of the FFNN network is presented in Table 3.

4 Experimental details and outcomes

The proposed method was executed on the machine running on the Window environment configured with i5, 8 GB RAM, 256 GB SSD and 4 GB Graphics support. The proposed method was implemented in Python 3 and used a synthesized shoplifting dataset for training and testing purposes. The details related to the dataset is given as follows:

4.1 Dataset collection

This work uses an annotated shoplifting dataset synthesized by ourselves for performing experiments. Here, the OpenPose algorithm makes annotations that generate body joints for the human present. The dataset contains overall 175 video clips representing two classes: normal and shoplifting. The clips representing usual actions like inspecting items, walking, talking, etc., are part of the normal class, while the clips representing human stealing acts like concealing items inside clothing and bag are part of the shoplifting class. Each clip is approximately 9–11 s in duration and is captured at 30 frames per second. The resolution of the recorded video clip is 640 × 480. Specific to our research, we took 128 clips from a synthesized shoplifting dataset that involves clear acts of normal and shoplifting. Out of 128, 68 clips represent human stealing actions (shoplifting), while the remaining 60 clips represent normal actions. In addition, we use a 75–25% training and testing split for performing training and validation tasks. Here, 96 clips are used in training and 32 clips are used in the validation process. Table 4 presents the annotated shoplifting dataset's distribution.

Table 3 FFNN configuration

Layers	Layer details	Activation
Input layer	4800	–
Hidden layer I	1024	Tanh
Hidden layer II	512	Tanh
Hidden layer III	256	Tanh
Output layer	2	Sigmoid

4.2 Performance measures

In this paper, the effectiveness of the proposed method is evaluated using Positive Predictive Value (PPV), True Positive Rate (TPR), False Positive Rate (FPR), Miss Rate (MR) and accuracy measures [7, 15, 19]. Here, PPV represents the truly positive instances out of total positive predicted instances, while TPR represents the truly positive instances out of all actual positive instances. In other contexts, PPV and TPR represent a measure of quality and quantity, respectively. Higher PPV means that a method returns more relevant instances than irrelevant ones and high TPR means that a method returns most of the relevant instances. PPV and FPR are calculated using Eqs. (9) and (10), respectively.

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

FPR represents incorrect positive instances out of all negative instances. A lower value in FPR is desirable, which means the method returns fewer false alarms for negative instances. It is given as

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{11}$$

Accuracy represents the correctly predicted instances out of all positive and negative instances. It is given as

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{12}$$

Miss rate identifies the total negative instances out of all positive and negative instances. It is given as

$$MR = \frac{FP + FN}{TP + FP + TN + FN} \tag{13}$$

Table 4 Annotated shoplifting dataset representation

Categories	Distribution	Clips
Shoplifting	Train	51
	Test	17
Normal	Train	45
	Test	15

where TP = The model correctly predicts the shoplifting class, TN = The model correctly predicts the normal class, FP = The model provides the wrong prediction of the normal class FN = The model wrongly predicts the shoplifting class.

4.3 Results and discussion

This subsection presents the experimental results of the proposed method evaluated on a synthesized shoplifting dataset. The method proposed by analyzing limited joints is trained on an FFNN classifier containing three hidden layers of different sizes. As presented in Table 5, the experimental outcomes show that the proposed method provides very few false-positive instances during the validation process and achieves a recognition rate of 96.87%. On the other hand, it offers good PPV and TPR with lower false alarms.

In addition to FFNN, the performance of the proposed method is also analyzed for some other classifiers too. The classifiers include Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF). Table 6 shows the comparative analysis of the proposed method over different classifiers. We found that the proposed method with LR classifier provides good accuracy with lower false alarms and good values of PPV and TPR during the validation process. This method in case of KNN classifier is tuned for $K = 5$ which offers a large false alarm rate with moderate values of accuracy, PPV and TPR. The SVM classifier is tuned for a linear kernel that scores good accuracy, PPV and FPR values, and less false positive instances for the proposed scheme. For Decision Tree (DT) classifier, the method scores moderate PPV, TPR, FPR and accuracy. For Random Forest (RF) classifier, the method scores the same accuracy as scored with Logistic Regression (LR). However, the method with RF classifier produces lower false-positive and remarkable false-negative instances than with LR classifier. Overall, our proposed method containing FFNN as a learner outperforms the others in terms of accuracy quantifies as 96.87%, and also delivers highest TPR, good PPV and lower FPR values.

To analyze the effectiveness of our proposed approach, classifiers used are also configured for three additional cases of experimentations. In the first case, the features are extracted for all 18 joints and used directly by a classifier. The second case extracts features for selective joints (i.e., 9) and uses them to perform classification. The third case extracts features for all 18 joints, uses PCA for feature reduction and then performs the classification on a reduced feature set. The last case is actually the proposal of this paper, experimented with the possible nearby hypothesis, to expose the novelty and efficiency of the proposal. The recognition rate of all the cases for different classifiers is presented in Fig. 5. We found that the first case acquires the lowest accuracy of 57.63% for the KNN classifier and the highest accuracy of 78.12% for the FFNN classifier. The second case offers up to 87.5% for DT classifier and the lowest accuracy of 68.12% accuracy for the KNN classifier. The third case achieves 93.75% accuracy for the FFNN classifier and the lowest accuracy of 75% for the KNN classifier. The proposed method or finalized case scores the highest accuracy, which is 96.87%, and the lowest accuracy of 81.25 for the KNN classifier. The proposed scheme with FFNN classification achieves 6.8%, 19.22%, 3.3%, 10.7% and 6.89% higher accuracies than with LR, KNN, SVM, DT and RF, respectively. It is also found that the proposed method outperforms other parallel alternatives in all three scenarios, as mentioned above.

In practice, it is still challenging to obtain accurate body poses due to missing portions of joints. This creates an imperfect result for automatic posture estimation. Here, the impact of missing body joints is examined to see the effectiveness of the video classification task. The recognition performance is evaluated for $n = 0, 1, 2, 3, 4$ and 5 missing body joints. To perform a wide range of experiments, n body joints are randomly chosen from selected nine body pairs, and then such body joints are skipped in the encoding process from different video sequences. Figure 6 presents the recognition accuracy calculated for different missing joints. On analysis, it has been observed that the performance of the proposed method decreases by increasing the number of missing joints. It achieves over 90% accuracy using only 7 out of 9 pairs and over 50% accuracy using 4 out of 9 pairs. The proposed method for SVM and FFNN can take care of up to two missing pairs without compromising any performance. Other hand, the proposed method cannot handle up to 5 missing joints more proficiently, causing degradation in accuracy. Finally, we infer that the proposed method can parse human behavior

Table 5 Experimental outcomes of proposed method over FFNN classification

Confusion matrix		Measures		Values	
15 (TP)	0 (FN)	Positive Predictive Value (PPV)		0.937	
		True Positive Rate (TPR)		1.00	
1 (FP)	16 (TN)	False Positive Rate (FPR)		0.058	
		Accuracy	96.87%	Miss rate	0.968

Table 6 Comparative analysis over different classifiers for synthesized shoplifting dataset

Classifier	TP	TN	FP	FN	PPV	TPR	FPR	MR (%)	Accuracy (%)
LR	14	15	2	1	0.875	0.933	0.117	9.3	90.62
KNN	12	14	3	3	0.800	0.800	0.176	18.75	81.25
SVM	14	16	1	1	0.933	0.933	0.058	6.25	93.75
DT	13	15	2	2	0.866	0.866	0.117	12.5	87.50
RF	13	16	1	2	0.928	0.866	0.058	9.37	90.62
Proposed method	15	16	1	0	0.937	1.000	0.058	3.12	96.87

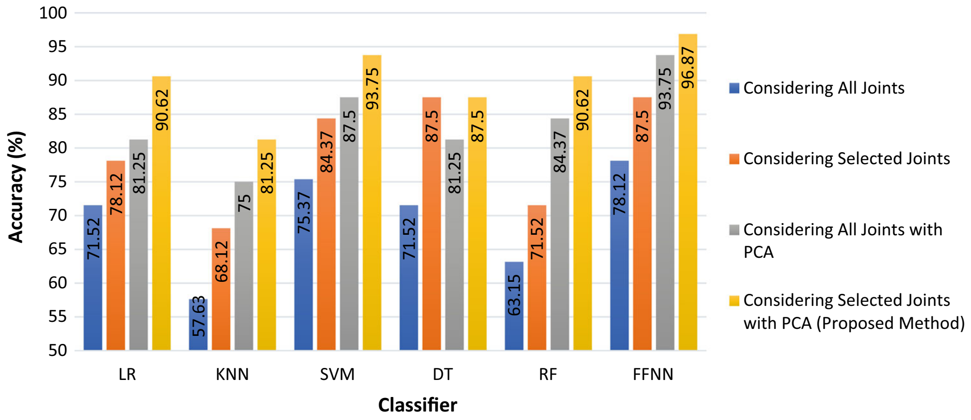


Fig. 5 Comparative analysis over different scenarios for synthesized shoplifting dataset

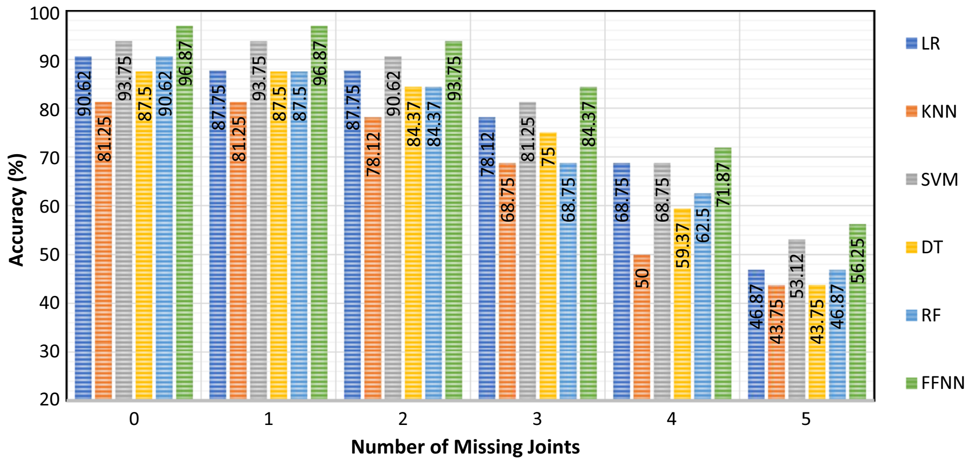


Fig. 6 Performance evaluation over different missing joints

with an arbitrary number of body joints as input and provide compelling results with few missing connections/ joints, up to 3 out of 9 joints.

Finally, the resulting instances of video clips evaluated over the proposed method are presented in Fig. 7. Here, the first and second rows depict the correctly classified instances of normal and shoplifting classes for self-occluded actions, respectively. Here, we infer that the proposed method worked upon the information is capable of categorizing human acts (i.e., normal and shoplifting) in an indoor environment very accurately.

The experiments as presented above analyzes the shopper behavior using proposed method over synthesized videos, which were recorded at the laboratory in a simulated environment. The next subsection presents the performance of the proposed method over real-store recorded videos. The detailed discussion is as follows.

4.4 Testing on real store camera footage

After analyzing human behavior in simulated environments, we also assessed the performance of proposed method on various real store camera footage, taken from the UCF crime dataset [20] and YouTube videos. A total of 20 videos are captured and used here to test our proposed method. These are distributed into two categories i.e.,

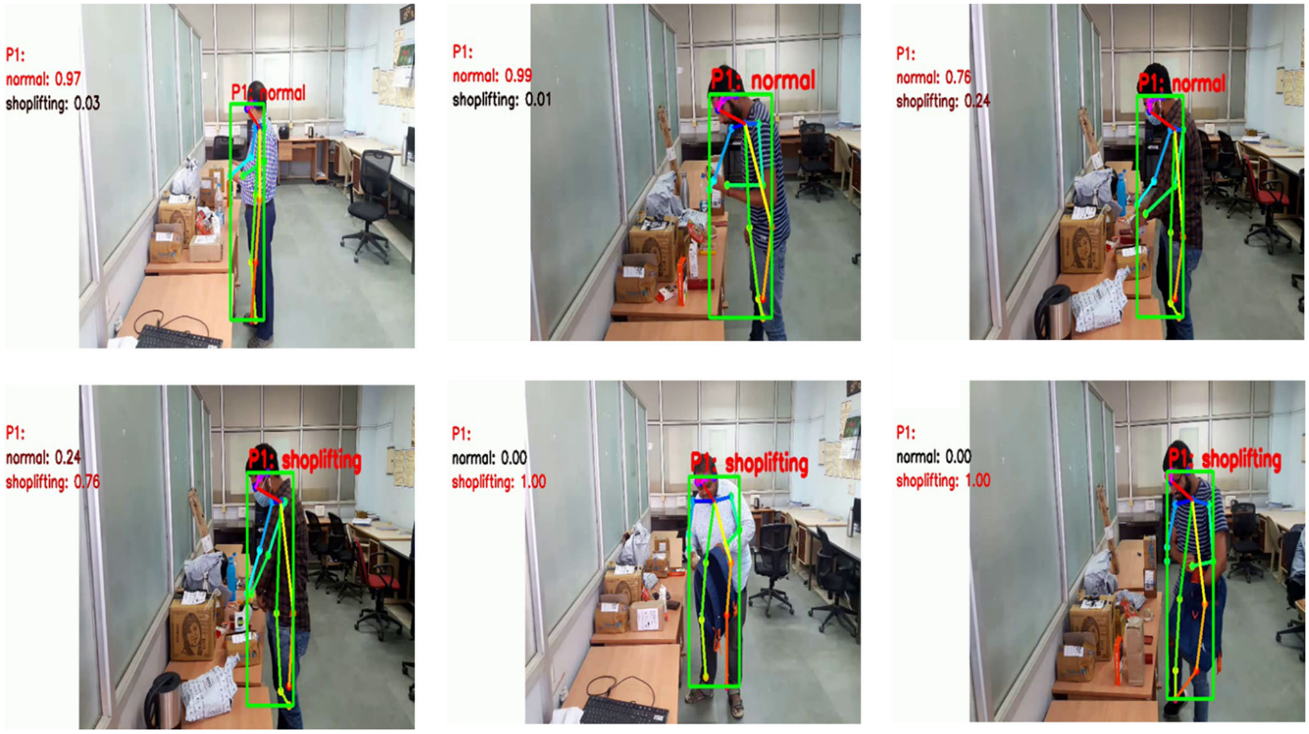


Fig. 7 Resulting instances evaluated over synthesized shoplifting dataset

Table 7 Comparative analysis over different classifiers for real store camera videos

Classifier	TP	TN	FP	FN	PPV	TPR	FPR	MR (%)	Accuracy (%)
LR	8	8	2	2	0.80	0.80	0.20	20	80
KNN	7	7	3	3	0.70	0.70	0.30	30	70
SVM	8	9	1	2	0.88	0.80	0.10	15	85
DT	8	7	3	2	0.72	0.80	0.30	25	75
RF	9	8	2	1	0.81	0.90	0.20	15	85
FFNN	10	8	2	0	0.83	1.00	0.20	10	90

normal and shoplifting, each containing ten videos. Table 7 presents the performance results of the proposed method for real-store recorded videos, while processing them for classification, using different classifiers. After analysis, we found that the proposed method gives large false positives with KNN and DT classifiers and false negatives with KNN classifier. However, this count of false positives and false negatives is very low in case SVM and FFNN classifiers. This means that skeleton-based processing with fine-grained details provides features more relevant for classification with SVM or FFNN. In terms of the evaluation scheme, the proposed method yields the highest PPV and lowest FPR values for the SVM classifier. As well as, it produces the highest TPR, lowest MR, and highest accuracy values for the FFNN classifier.

It has already been seen that it is more convenient to process selected body joints to analyze shopper's behavior than to process all the body joints. Therefore, the real-store recorded videos are also examined for the two different scenarios. The first case predicts shopper behavior by using selective body joints with a classification technique. The other or second case is our proposal, which uses selected body joint pairs, a feature reduction technique, and a classification technique to analyze the individual's behavior. Figure 8 shows the comparative analysis of the proposed method in both the cases. For the KNN classifier, both cases score almost identical outcomes. However, the first case yields the lowest accuracy for the DT classifier and the highest accuracy for the FFNN classifier. Other side, the second case achieves the lowest accuracy for KNN classifier and highest accuracy for the FFNN classifier. From this, we confer that the proposed method with the FFNN classifier outperforms all other scenario's performed with different classifiers.

In Fig. 9, we have shown activity detection with real store recorded videos. These recorded videos show normal

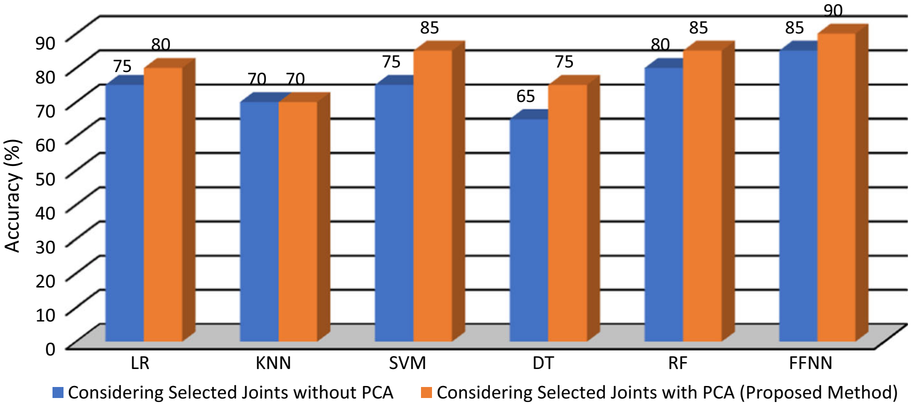


Fig. 8 Comparative analysis over different scenarios for real store camera video

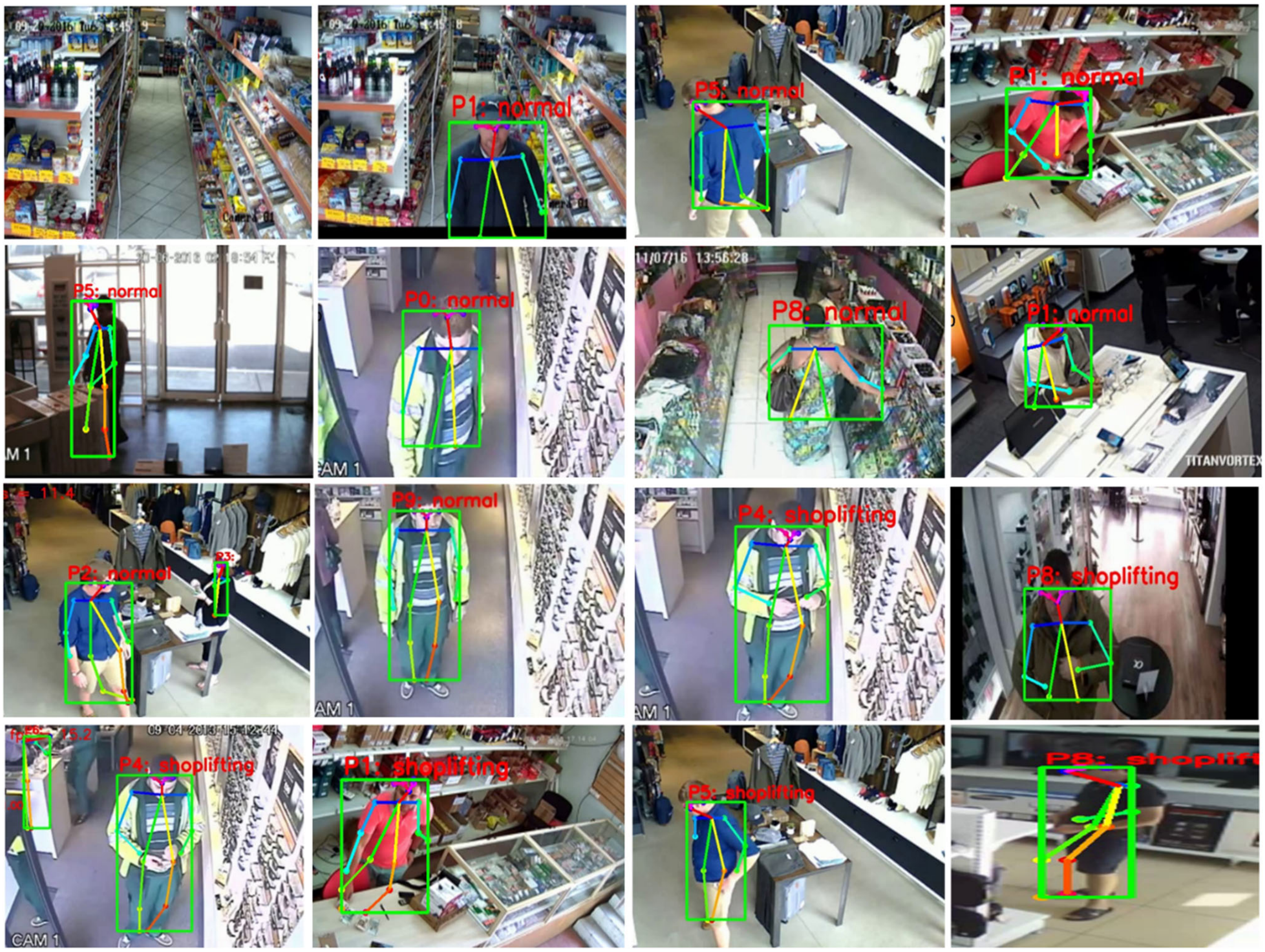


Fig. 9 Resulting instances evaluated on various real store camera footage

and shoplifting-related human behavior at shops and stores. In addition, the recorded sequences of the real-store footage include some variations such as illumination variation, pose variation, occlusion, etc. After analysis, we found that our proposed method shows encouraging test results on real-store recorded videos and classifies human behavior in real-time with an accuracy of 90%.

4.5 Comparisons with existing methods

This section compares the recognition rate of our proposed method with existing state-of-the-art methods. Here, Fig. 10 presents a comparative analysis between the proposed method and other existing methods. Analysis tells that the proposed method evaluated over synthesized

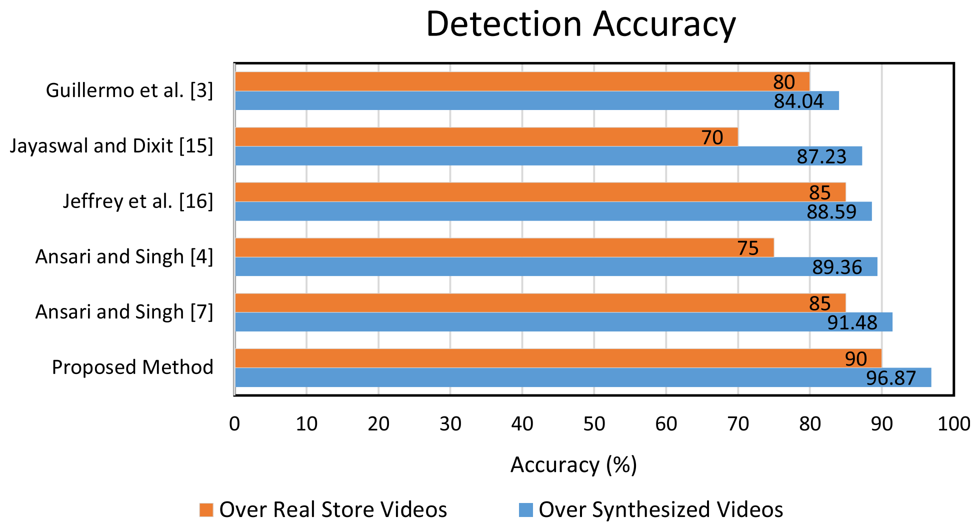


Fig. 10 Comparisons with existing state-of-the-art methods

videos is more accurate up to 15.26% than Method [3], 11.05% than Method [15], 9.34% than Method [16], 8.40% than Method [4], and 5.89% than Method [7]. Similarly, in the case of processing real store recorded videos, the proposed method scores more accurate outcomes up to 12.5% than Method [3], 28.57% than Method [15], 5.88% than Method [16], 20% than Method [4], and 5.88% than Method [7]. The discussion related to existing methods has already been covered in the literature part of this paper. Finally, we infer that the proposed method has better performance than other available existing methods.

5 Conclusion

This paper couples the pose estimation and action recognition processes in a unified framework. This proposed framework is focused on detecting shoplifting actions using 2D postural data encoded over various parameter spaces. Encoding poses in the inputted skeleton data before extracting the features helps a lot in configuring the occluded actions. The pertinent features like normalized joints ($Xnorm$), joint motion (V), joint angle (θ) and distances (D) are calculated from the posture encoded data. Here, the feature reduction process plays a dominant role in making the features more intuitive, indicating the effectiveness of the proposed method. Finally, classification is used to classify the human acts based on reduced feature sets, and a recognition rate of 96.87% is achieved for the fully connected FFNN classifier over synthesized videos. We also found that the proposed model can detect human stealing actions even in partially occluded environments and can efficiently handle missing body joints. In addition, the proposed method validated over real-store recorded videos gives promising results up to 90% accuracy. The performance of the proposed method is found to be exceptional compared to the other methods available in the literature. This showcases the superiority of this method and its applicability in real-time surveillance problems.

Funding None.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Deccan Herald (2019) Retail chains battle shoplifting. https://www.deccanherald.com/metrolife/retail-chains-battle-shoplifting-722520.html. Accessed 22 Oct 2021
2. Arroyo R, Yebes JJ, Bergasa LM, Daza IG, Almazán J (2015) Expert video-surveillance system for real-time detection of suspicious behaviours in shopping malls. Expert Syst Appl 42(21):7991–8005
3. Martínez-Mascorro GA, Abreu-Pederzini JR, Ortiz-Bayliss JC, Garcia-Collantes A, Terashima-Marín H (2021) Criminal intention detection at early stages of shoplifting cases by using 3D convolutional neural networks. Computation 9(2):24
4. Ansari MA, Singh DK (2022) An expert eye for identifying shoplifters in mega stores. In: International conference on innovative computing and communications, pp 107–115
5. Beddiar DR, Nini B, Sabokrou M, Hadid A (2020) Vision-based human activity recognition: a survey. Multimed Tools Appl 79(41):30509–30555
6. Khan NS, Ghani MS (2021) A survey of deep learning based models for human activity recognition. Wirel Pers Commun 120(2):1593–1635

7. Ansari MA, Singh DK (2022) An expert video surveillance system to identify and mitigate shoplifting in megastores. Multimed Tools Appl 81(16):22497–22525
8. Munea TL, Jembre YZ, Weldegebriel HT, Chen L, Huang C, Yang C (2020) The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation. IEEE Access 8:133330–133348
9. da Silva MV, Marana AN (2020) Human action recognition in videos based on spatiotemporal features and bag-of-poses. Appl Soft Comput 95:106513
10. Dwivedi N, Singh DK, Kushwaha DS (2020) Orientation invariant skeleton feature (OISF): a new feature for human activity recognition. Multimed Tools Appl 79(29):21037–21072
11. Li C, Tong R, Tang M (2018) Modelling human body pose for action recognition using deep neural networks. Arab J Sci Eng 43(12):7777–7788
12. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
13. Sanal Kumar KP, Bhavani R (2020) Human activity recognition in egocentric video using HOG, GiST and color features. Multimed Tools Appl 79(5):3543–3559
14. Smagulova K, James AP (2019) A survey on LSTM memristive neural network architectures and applications. European Phys J Spec Top 228(10):2313–2324
15. Jayaswal R, Dixit M (2021) A framework for anomaly classification using deep transfer learning approach. Rev d'Intelligence Artif 35(3):255–263
16. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
17. Noori FM, Wallace B, Uddin M, Torresen J (2019, June) A robust human activity recognition approach using OpenPose, motion features, and deep recurrent neural network. In: Scandinavian conference on image analysis, pp 299–310
18. Wang R, Wu X (2019) Combining multiple deep cues for action recognition. Multimed Tools Appl 78(8):9933–9950
19. Singh DK (2018, July) Human action recognition in video. In: International conference on advanced informatics for computing research, pp 54–66
20. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6479–6488

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.