Article

Early Detection of Collective or Individual Theft Attempts Using Long-term Recurrent Convolutional Networks

Khaled Hoshme [1]

[1] Department of Software Engineering and Information Systems, Al-Baath University, Homs, Syria;

Abstract: Theft crimes cause many losses to many facilities and companies around the world, this leads to a considerable number of risks, and despite the spread of a large number of surveillance cameras, and large surveillance teams that track the movement that takes place within the store, the planning of many thieves can be Able to carry out the theft process without being noticed by human observers, as the human sight has movement limits, where the human observer can overlook one of the screens that records the actual movement at a moment, and therefore the theft can take place at this moment without Pay attention to it, and pre-planning the robbery can lead to the inability of human eyes to detect these attempts to carry out various theft operations. We proposed a model based on convolutional neural networks and recurrent neural networks to study the behavior and body language of shoppers within stores, where the proposed system can detect individual theft attempts or collective attempts to carry out the theft process. While other methods identify the crime itself, we instead model suspicious behavior - behavior that may occur before the accretion stage of crime - by exposing minute parts of the video with a high probability of containing the crime of shoplifting. Movement, which can lead to theft, the proposed neural structure was trained on a large number of visual clips that include attempts to steal according to a specific methodology. Through the proposed system, we reached an accuracy of 93 percent and a confidence coefficient of 93 percent.

Keywords: Theft; Shoplifting; Pre-crime Behavior Method; Convolutional Neural Network; LSTM, Suspicious Behavior;

1. Introduction

The theft operations on the store and its planning are one of the most complex matters, as the human observers inside the control center have to follow all the movements of the people who are shopping inside the store in real-time, and in most cases, the human observer cannot recognize the thefts, and this is due to the uncertainty of That the movements made by thieves are movements that can lead to theft, and therefore one of the other reasons for the inability of the human observer to identify the theft operations is the prior planning of a group of thieves to carry them out, the aim of which is to distract the observers and thus the theft process takes place without the ability of a human observer to identify on her [1].

Several studies have emerged that provide a typical architecture for an artificial intelligence model capable of analyzing motion, whether to pre-identify the possibility of theft or other models to determine the actual theft. In this section, we will mention many studies and methodologies that were taken to develop intelligent motion recognition systems, and analyze and verify them.

The study [2] depends on early movement analysis to detect anomalies in movement and to identify the intent to commit theft. The study relied on the use of a dataset that includes several videos of theft operations, which were recorded through surveillance cameras. Theft includes several periods, starting from the moment the thief appeared in the video, the period during which the movement was strange or abnormal (which is the movement that was focused on discovering the intention to commit the theft), as well as the period during which the theft was carried out, and the period during which the theft took place. Movement returned to normal.

The study [2], relied on determining the beginning of the time when the suspicious movement started and the time at which the suspicious movement ended.

In addition to the natural videos that the study relied on, and therefore the study relied on two types, the first type includes videos of the movement that precedes the theft and natural videos.

Figure 1 Video segmentation by using the moments obtained from the Pre-Crime Behavior Segment (PCB) method [2].

In addition, the study relied on three-dimensional convolutional neural networks (3DCNN) to study the movement and identify the characteristics in those videos to build a mathematical model, capable of analyzing strange movements, the following figure shows the structure of the three-dimensional neural network that was used in the study [2].

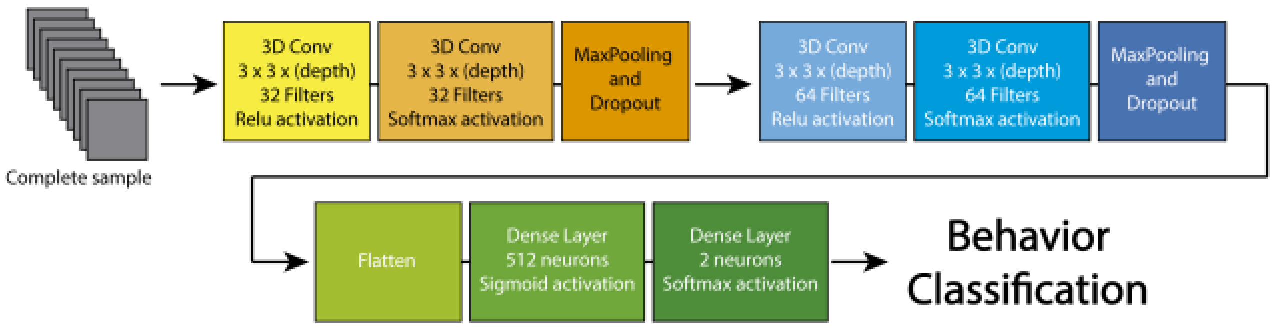

Figure 2 Architecture of the DL Model used in [2]

The study [2] reached an accuracy of approximately 85 percent in classifying the videos, and the study was able, in most cases, to identify the suspicious movement that precedes the occurrence of the crime.

The [3] study, propose a framework to identify abnormal behaviors through deep-learning-based detection of non-semantic-level human action components segmented with a window size of several seconds (e.g., walking, standing, and watching) and performing sequence analyses of the detected action components to infer behavior intentions. Then, tested the applicability of the framework to the specific scenario of shoplifting, one of the most common crimes. Analysis of actual incident data confirmed that shoplifting intentions could be effectively gauged based on distinct action sequence features, and the intention inference results are continuously updated with the accumulated series of detected actions during the course of the input video stream. The results of this study can help enhance the ability of intelligent surveillance systems by providing a new means for monitoring abnormal behaviors and deeply understanding the underlying intentions.

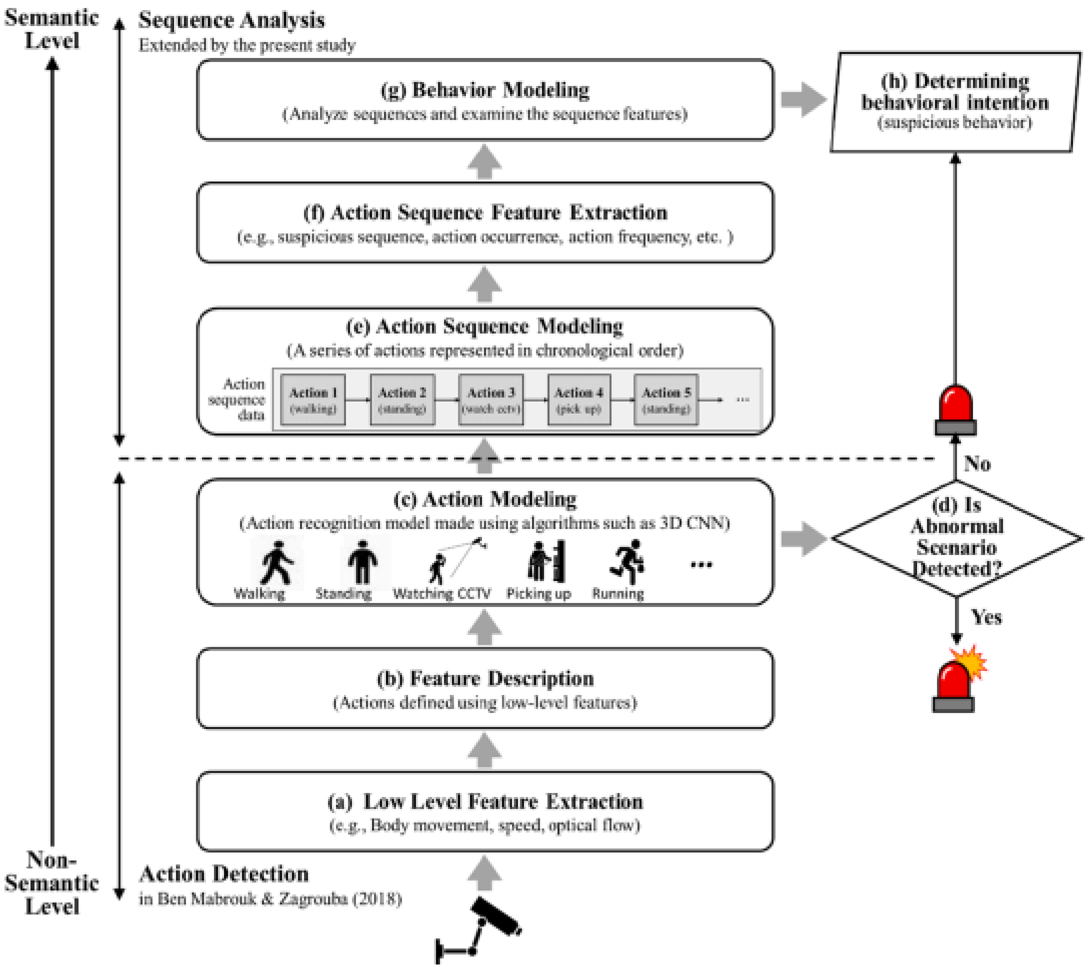

Figure 3 . Framework for identifying abnormal behaviors and inferring the intentions underlying specific behaviors used in [3]

The study [3] mainly depends on the discovery of strange movement through a study that follows a set of movements that can be performed by a specific person, and therefore the study depends on the sequence of execution of a set of movements for a specific person can perform or be classified as a strange movement, and therefore the study [3], classifies the movement of one person and is unable to study the intelligence of a group of people to carry out the theft.

The following figure illustrates the methodology used in the study [3], where each person is initially extracted independently and each person's movement is studied.

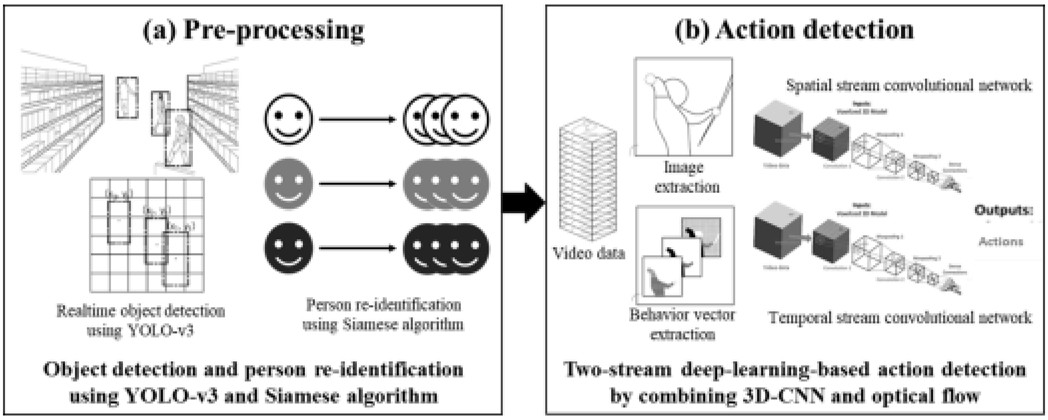

Figure 4 Overview of action detection process using 3D-CNN algorithm [3].

The study used [3], YOLO-v3 to identify the person in the monitored area, after which each person is passed to the 3D neural network 3DCNN.

2. Methodology

2.1 Description of The Dataset

In our proposed system, based on the availability of several videos that include thefts, we worked on using the concepts that were built and used in the study [2], but with great improvement in dealing with the dataset used, and moments in time, and the aim of this is the ability to access The accuracy is higher than that obtained in the study [2], and thus in this work, we use the UCF-Crime dataset [9] to analyze suspicious behavior during the accumulation of shoplifting crime. The dataset consists of 1900 real-world observational videos and provides about 129 hours of video clips. Videos are not normalized and are displayed at a resolution of 320 x 240 pixels. The dataset includes scenarios from multiple people and locations, grouped into 13 categories such as “abuse,” “burglary,” and “explosion,” among others. We extracted samples used in this investigation from the 'store robbery' and 'normal' categories from the UCF-Crime data set.

As for the videos of the thefts, we divided those videos into three different periods, as follows:

- The moment the thief appeared, which was his natural movement.
- The moment the suspicious movement began for the thief.
- The moment the thief made the theft.
- The moment the theft was completed.
- The moment the movement returned to normal.

According to this division, we will extract periods for both the normal movement and the abnormal movement from the videos of the theft, as this case was taken care of, to increase the neural network to identify the change in movement that led to the transformation of the movement from the normal movement to the suspicious movement, and after So returning to the normal movement again, this process will contribute to helping the neural network to discover the suspicious movement within a group of people who are within the same monitored area, and thus the ability to identify the suspicious movement.

Therefore, the methodology for collecting the dataset relied on taking advantage of the videos of the theft process to extract the moments when the movement was normal and the moments when the movement was abnormal, and as we mentioned, this process will help the neural network to accurately identify the strange movement and the sudden change that occurred in The normal movement turned into an abnormal movement.

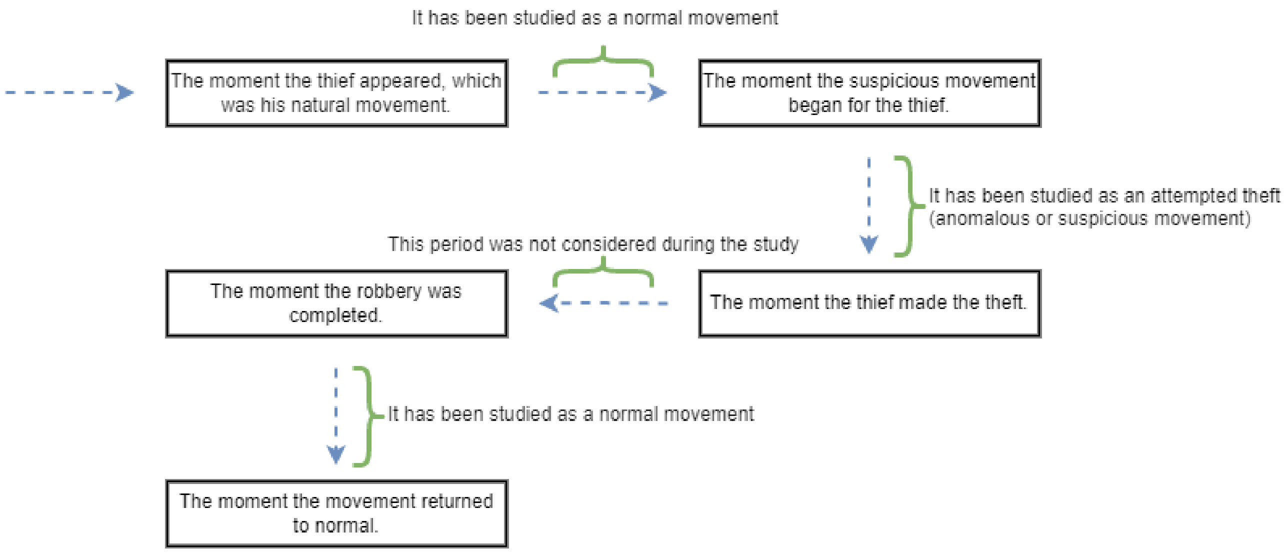

Figure 5 Segmentation and collection of visual moments from videos that include cases of theft

At the same time, several videos of the natural videos category were taken advantage of, as all the videos that were recorded through surveillance cameras were selected, within stores and shopping centers.

We have also increased the accuracy by collecting additional videos of the various thefts, as YouTube was used to search for clips of different thefts using the Russian, Persian, and French languages.

The number of theft videos that were used to train the neural network is 104, and the number of videos representing natural motion is 109.

Each period for a specific case was divided into several parts to increase the accuracy of the neural network in identifying the details of the movement, as the movement was divided into 400 frames each time. The neural network learns the details of suspicious movement and natural movement more accurately.

The number of parts that were reached after studying the movement every 400 symbols reached 289 video clips representing the suspicious movement (the movement that precedes the theft), and 320 video clips representing the natural movement.

```
Normal_Videos_001_x264.mp4  Normal  0  540  -1  -1
Normal_Videos_002_x264.mp4  Normal  0  400  -1  -1
Normal_Videos_002_x264.mp4  Normal  400  800  -1  -1
Normal_Videos_002_x264.mp4  Normal  800  1200  -1  -1
Normal_Videos_002_x264.mp4  Normal  1200  1600  -1  -1
Normal_Videos_003_x264.mp4  Normal  0  400  -1  -1
Normal_Videos_003_x264.mp4  Normal  400  800  -1  -1
Normal_Videos_003_x264.mp4  Normal  800  1200  -1  -1
Normal_Videos_003_x264.mp4  Normal  1200  1600  -1  -1
Normal_Videos_003_x264.mp4  Normal  1600  2000  -1  -1
Normal_Videos_003_x264.mp4  Normal  2000  2400  -1  -1
Normal_Videos_003_x264.mp4  Normal  2400  2800  -1  -1
```

Figure 6 Video clips that include natural movement, and the numbers of frames that begin and end with natural movement.

```
Shoplifting001_x264.mp4  Shoplifting  0  420  -1  -1
Shoplifting001_x264.mp4  Shoplifting  420  870  -1  -1
Shoplifting001_x264.mp4  Shoplifting  870  1230  -1  -1
Shoplifting003_x264.mp4  Shoplifting  6690  6900  -1  -1
Shoplifting004_x264.mp4  Shoplifting  2100  2600  -1  -1
Shoplifting004_x264.mp4  Shoplifting  2600  3000  -1  -1
Shoplifting005_x264.mp4  Shoplifting  0  750  -1  -1
Shoplifting006_x264.mp4  Shoplifting  270  900  -1  -1
Shoplifting006_x264.mp4  Shoplifting  900  1710  -1  -1
```

Figure 7 Video clips with suspicious movement, and the frames that begin and end with the suspicious movement.

2.2 Background Removal:

Since the videos included in the dataset are collected in many shopping centers, and because the distribution of goods in those centers can vary from one center to another, and because we only study the movement, so we only care about the movement of people, at this point, we focus On the moving people only, and therefore, we removed the background. The methodology we followed to remove the background and identify the edges of the people who move within the video depends on the study of changes in the values of the changing pixels between two successive frames. In contrast, the methodology depends on subtracting the values of the pixels between two Consecutive frames in absolute value. Thus until reaching 160 frames, subtraction operations are performed between the pixel values for every two successive frames, the proposed methodology helped to extract only the changing movement within the video, and thus to identify the movement of people only.

2.3 Shadow Removal:

Since the system studies the movement of people and because the lighting in the stores can be distributed differently, and here and for the importance of focusing on the movement of people only and removing all other information that can affect the training process, we suggested removing the shadow caused by the movement of people according to the position of the lighting in the stores, Thus, we suggested, after completing the background removal stage, to remove a few small pixel values whose value does not exceed 10, and here we can through this stage remove the small pixel values, which can represent the shadow of the moving people in the videos included in the dataset.

2.4 Remove Unimportant Details:

Since the visuals included in the dataset may consist of some unimportant details, which can affect the process of removing the background and focusing on the movement of people (such as several moving objects other than people), and thus avoid those moving objects contained in the visuals included in the dataset, use Gaussian blur "Gaussian blur is an image processing technique that results from blurring an image using a Gaussian function. Gaussian blur is widely used in graphics software, and usually reduces image noise as well as unwanted detail."

Also, the use of Gaussian blur greatly helped to focus on the edges of the moving people within the visual clips, without entering the rest of the details such as eyes, mouth, and other details within the motion analysis.

2.5 Data Augmentation:

Since the system analyzes the movement by detecting the intent of theft within the store, and since we need to generalize the ability of the neural network to be able to analyze the movement of people in the monitored area, we suggested using Data Augmentation to generate a more significant number of videos that include clones of the same basic videos but According to a different direction and angle of inclination, and the goal is the ability of the neural network to analyze the movement, according to many different causes, and therefore we proposed to generate additional videos from the same basic videos, but with a change in the direction of movement horizontally, and also with the use of a tilt angle of 30 degrees, so we are In this case we have generated many additional cases of people moving with the change of direction and angle of inclination.

The goal of using the 30-degree tilt angle for the videos generated using the concept of Data Augmentation is to simulate the position of the surveillance camera and to provide the greatest power to the neural network during the training phase to generalize its findings.

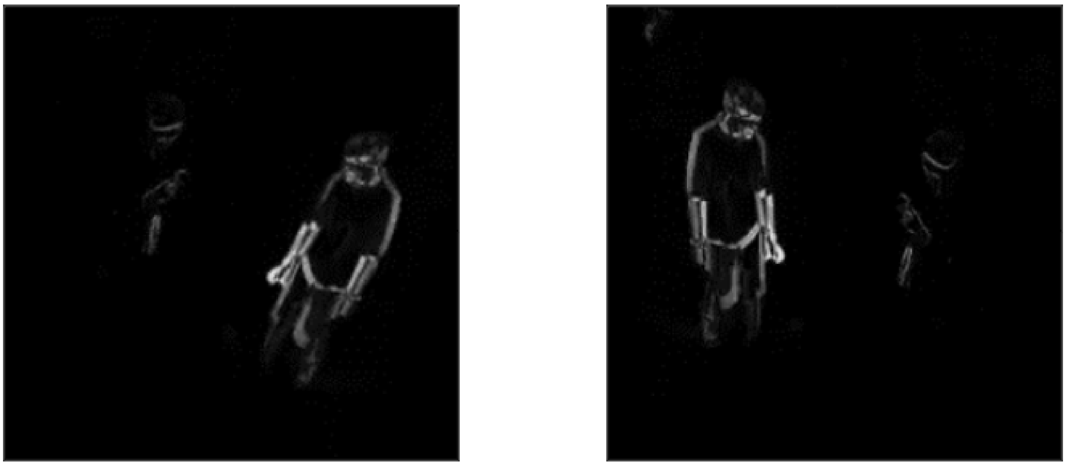

Figure 8 Video Data Augmentation Example.

2.6 Frame Metrics:

During the development phase, the encoding of the visual clips was converted to grayscale, in addition to the size of the frames used during the development phase (90 and 90) for both length and width, with the use of a sequence of 160 consecutive frames, and therefore the input of the neural network is as follows (160, 90, 90, 1), which represent (the sequence of a consecutive number of frames, frame length, frame width, coloring pattern).

The dataset was divided into a section for training and a section for testing, where 10% of the videos included in the dataset were approved for testing, and 90% for training.

3. Long-term Recurrent Convolutional Networks

The neural network that was used during the development stage includes two stages, the first includes extracting properties from the frames contained in the visual clips and passing those characteristics to the recurrent neural network layer and the long-term memory LSTM. The recurrent neural network and long-term memory are in the form of a one-dimensional array, where the long-term memory and the recurrent neural network study the relationship of the sequence of pixel values of each frame of the sequence of frames contained in the visual clips contained in the dataset used [10].

In this part, we will review the structure of the neural network that was used to study movement, which is a mixture of the convolutional neural network and the recurrent neural network, where we will review the proposed structure gradually.

3.1 Convolutional Neural Network:

The convolutional neural network used to study the characteristics of each frame of the sequence of frames contained in each visible segment of the dataset consists of several layers in addition to determining the characteristics of each of the layers that are used, the following figure shows the structure of the neural network The convolution that was used to identify the properties contained in the frame.

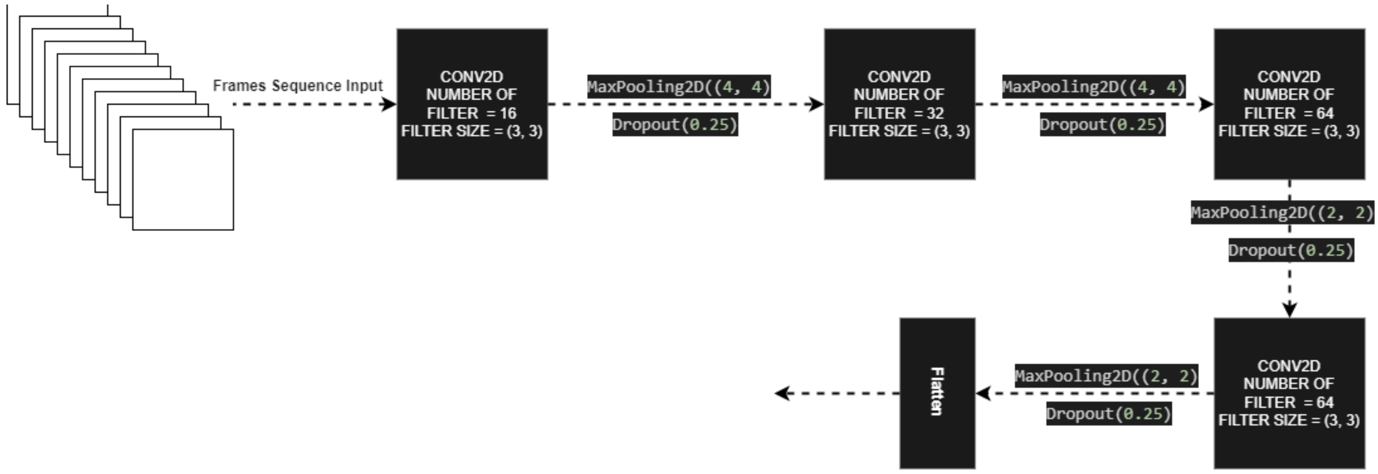

Figure 9** **The architecture of a convolutional neural network that was used to extract properties from the frames.

3.2 Long Short-term Memory:

We proposed the use of long-term memory to study the sequence of pixel values of the frames contained in each video clip. Gates, through which long time sequences can be studied, and since we use a sequence of frames up to 160 to analyze the movement resulting from the sequence of that number of frames, we suggested the use of long-term memory in studying the characteristics that were extracted from the CNN layers and thus trying to link those characteristics to each other To identify and sort the movement.

The following figure illustrates the gates that comprise the LSTM:

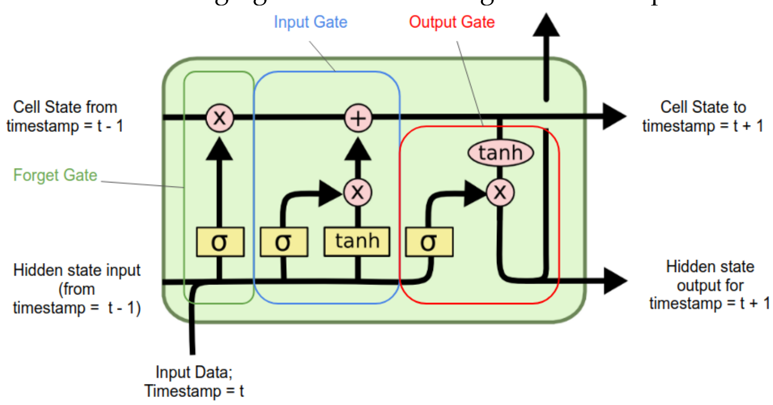

Figure 10 A single LSTM Cell [11]

Forget Gate: Its task is to determine which information is ignored by the unit.

Input Gate: Define output values to update the memory state.

Output Gate: Determine what is output according to the input unit and memory.

3.3 Time Distributed:

The architecture of a convolutional neural network can extract the properties of one image at a time (one frame at a time), but here we have a sequence of frames, and therefore we need a way through which we can pass that sequence from frames to the two-dimensional convolutional neural network.

We proposed the use of Time Distributed so that the convolutional neural network can receive more than one input to it (more than one image at the same time), and thus we can pass the 160-frame visual segment to the two-dimensional convolutional neural network. A convolutional neural network extracts the properties of each frame independently from the next.

We need to use Time Distributed so that the characteristics that distinguish each frame can be extracted so that the long-term memory can study the characteristics that were extracted from each frame.

The following method is based on linking the outputs (characteristics) extracted from each frame independently and passing them to long-term memory to study the sequence of those characteristics (the characteristics of each frame of the sequence of frames that make up the studied video).

The following figure shows the working mechanism of Time Distributed so that we can pass more than one frame (image) at the same time to the convolutional neural network, and then the properties of each frame are transformed into a single-distant array, and the output is transmitted to the long-term memory LSTM, which studies the sequence of those The properties to link those sequences into how the movement type is defined and sorted, and finally, the results are passed to the Dense layers to link the results obtained and extract the corresponding mathematical model.

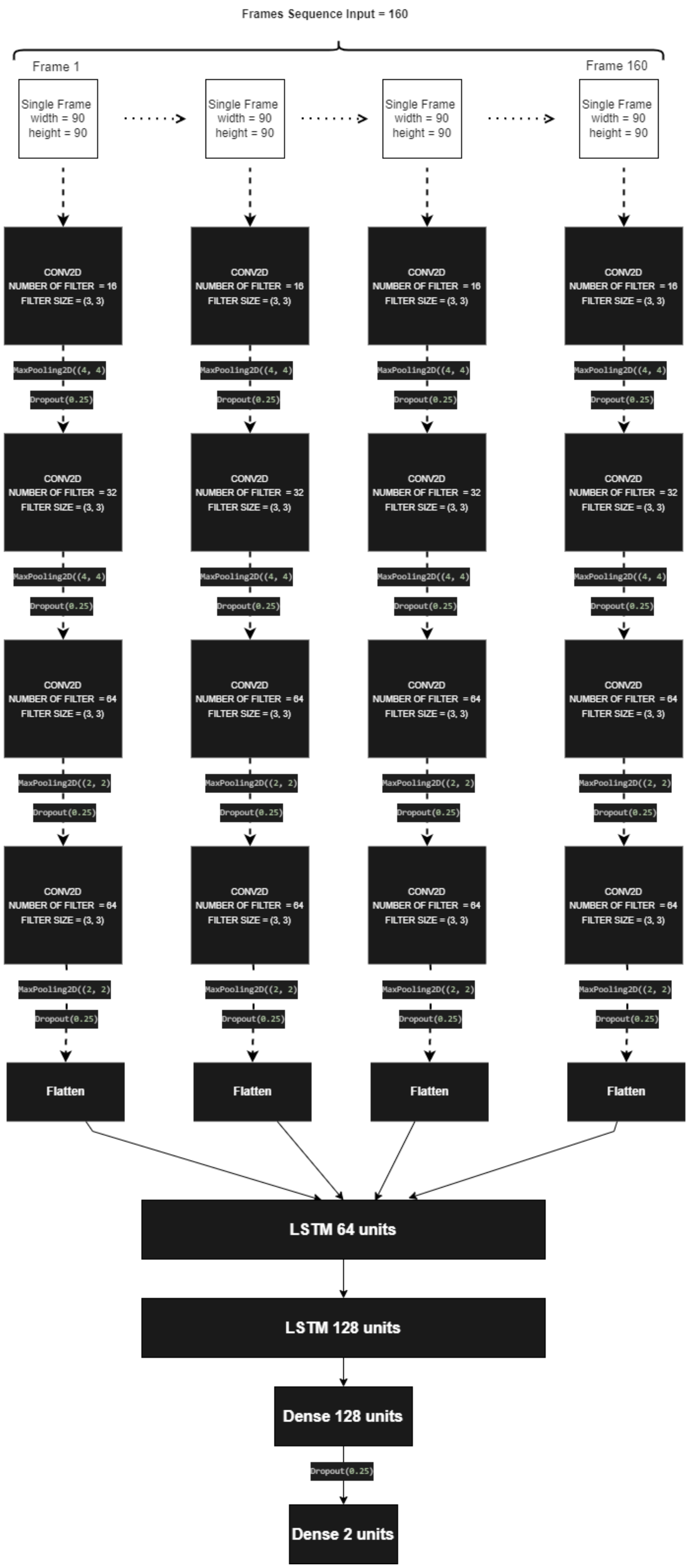
Frames Sequence Input = 160
Frame 1
Frame 160
Single Frame
width = 90
height = 90
CONV2D
NUMBER OF FILTER = 16
FILTER SIZE = (3, 3)
MaxPooling2D((4, 4)
Dropout(0.25)
CONV2D
NUMBER OF FILTER = 32
FILTER SIZE = (3, 3)
MaxPooling2D((4, 4)
Dropout(0.25)
CONV2D
NUMBER OF FILTER = 64
FILTER SIZE = (3, 3)
MaxPooling2D((2, 2)
Dropout(0.25)
CONV2D
NUMBER OF FILTER = 64
FILTER SIZE = (3, 3)
MaxPooling2D((2, 2)
Dropout(0.25)
Flatten
LSTM 64 units
LSTM 128 units
Dense 128 units
Dropout(0.25)
Dense 2 units

Figure 11 The structure of the neural network (linking spatio-temporal properties) used in the study.

The proposed neural structure linking spatial characteristics with time helped extract the most essential characteristics, which can be considered a criterion for determining the possibility of a case of theft. The convolutional neural network to extract the spatial characteristics from the frames and then move to the use of the recurrent neural network to link those spatial characteristics with time helped to extract the general characteristics and focus on them during the process of analyzing and classifying the sequence of frames.

3.4. Result:

As a critical measure for analyzing the results, we took into account accuracy (Equation (1)). It takes into account correct results—true positive (TP) as well as true negative (TN)—on the total number of samples evaluated (FP and FN represent false positives and false negatives, respectively).

Since accuracy shows the overall performance of the model, we supplement its information by using two additional measures to adequately analyze the results: precision (Equation (2)) and Recall (Equation (3)). Precision indicates the proportion of samples classified as suspicious that are suspicious - a model with a precision of 1.0 does not produce an FP. Recall indicates the percentage of actual suspicious samples that have been correctly classified by the system.

At the same time, the Figure 15 shows the loss during the training process and shows the actual decrease in the loss during the training stages that the neural network went through.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

We will review several schematics that illustrate the training process of the proposed neural network so that each scheme expresses a specific mathematical equation, or shows and adopts the increase in the accuracy and reliability of the proposed model.

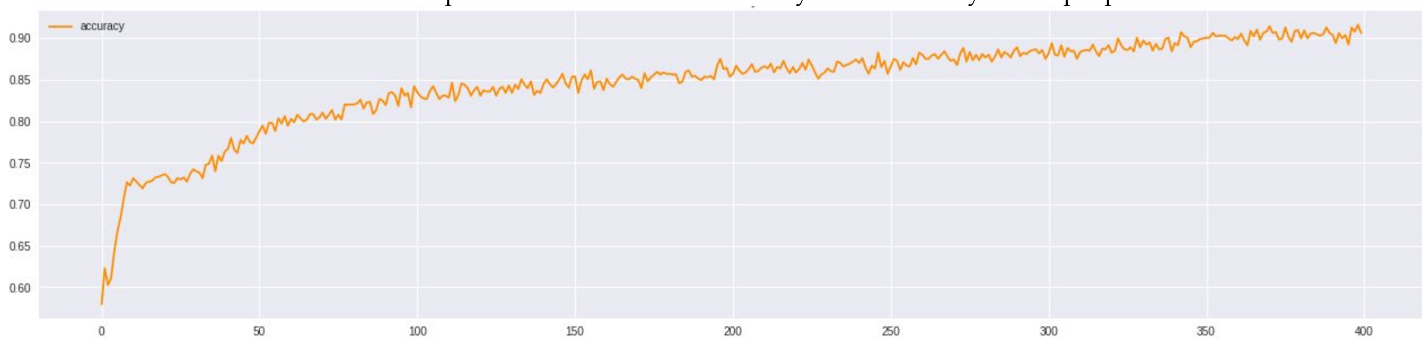

Figure 12 Diagram showing the increase in accuracy during the training phase.

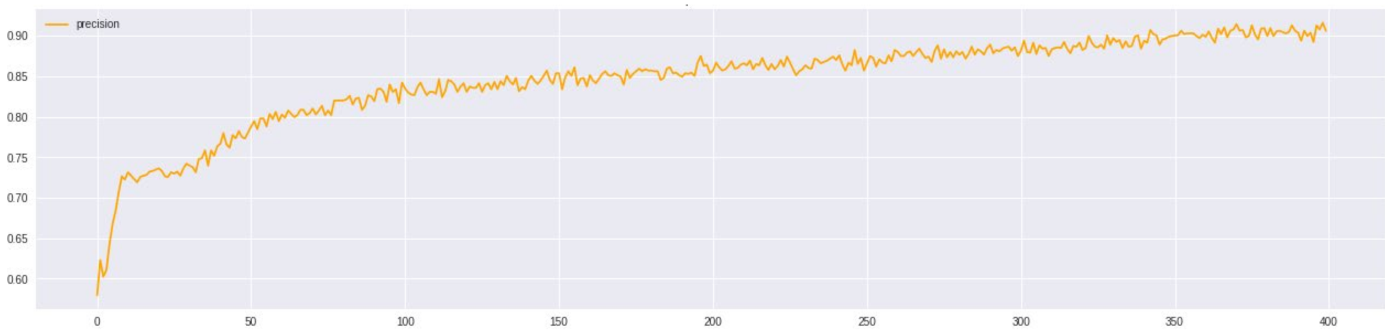

Figure 13 Diagram showing the increase in precision during the training phase.

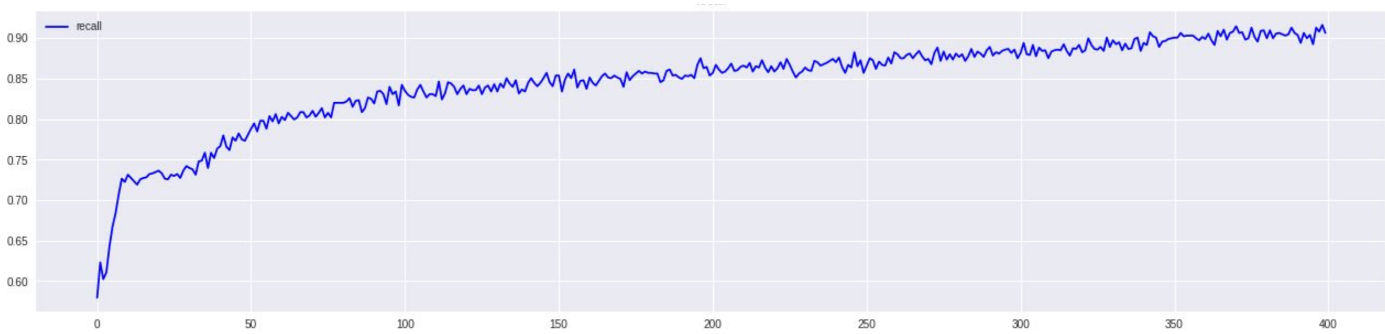

Figure 14 Diagram showing the increase in recall during the training phase.

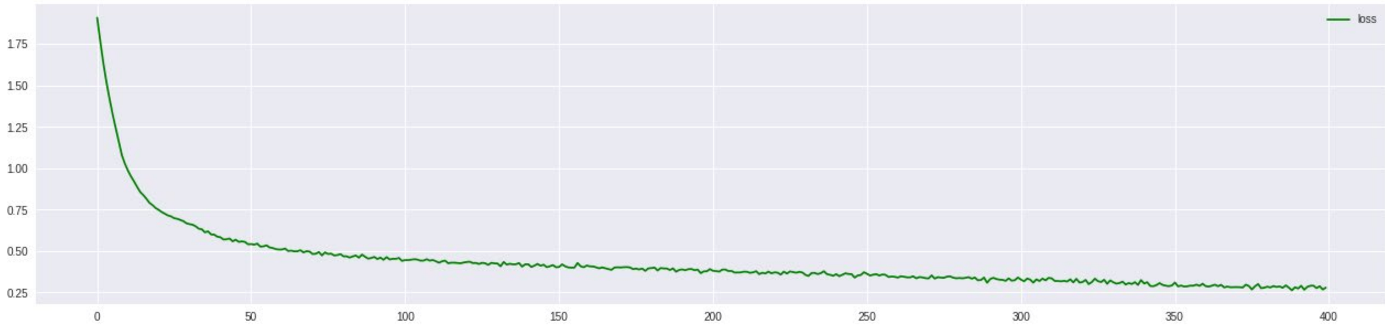

Figure 15 Chart showing the decreasing value of the loss during the training phase.

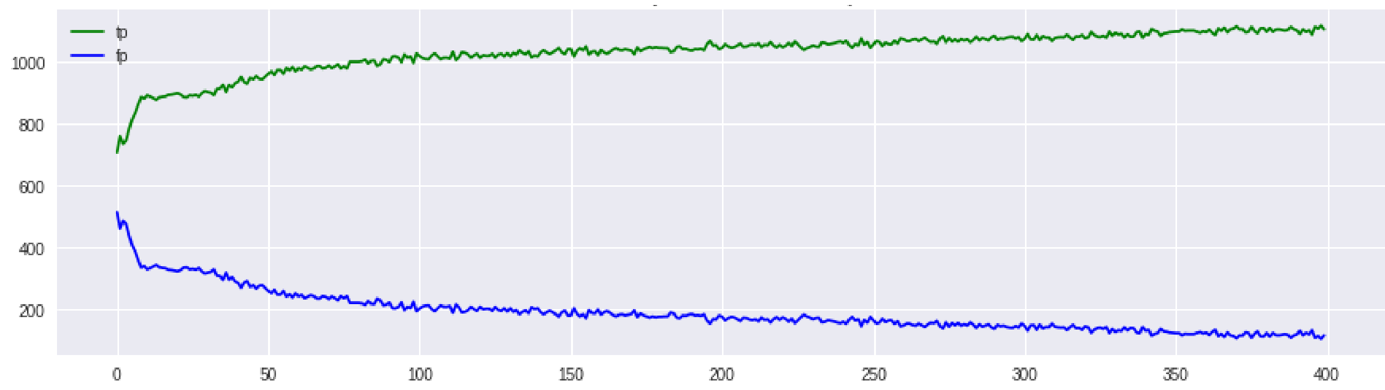

Figure 16 Changing TP and FP Values during the training phase.

Accuracy is a very commonly used metric, even in the everyday life. In opposite to that, the AUC is used only when it's about classification problems with probabilities in order to analyze the prediction more deeply. Because of that, accuracy is understandable and intuitive even to a non-technical person.

The Figure 17 reviews the accuracy and AUC diagrams, showing in-depth the increase in system accuracy during the training process.

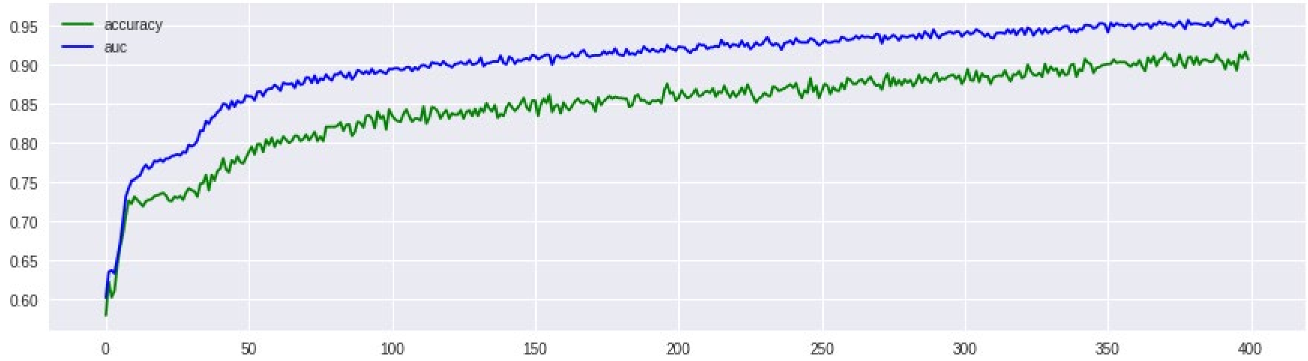

Figure 17 Accuracy and AUC during the training phase.

To clarify the results that were reached after completing the training, a confusion matrix was used on the test data, where the figure shows the results reached, 0 represents suspicious movement, and 1 represents normal movement, and as shown in the confusion matrix, the model was able to classify all cases which include suspicious movement and was able to classify 80% of the test data for normal movement as normal movement, while the remaining 20% of normal movement was classified as suspicious.

The confusion matrix shows the accuracy of the model in detecting suspicious movement that precedes cases of theft, as well as identifying the natural movement, although in some cases it classified the natural movement as suspicious, the model was able to classify all suspicious movements successfully.

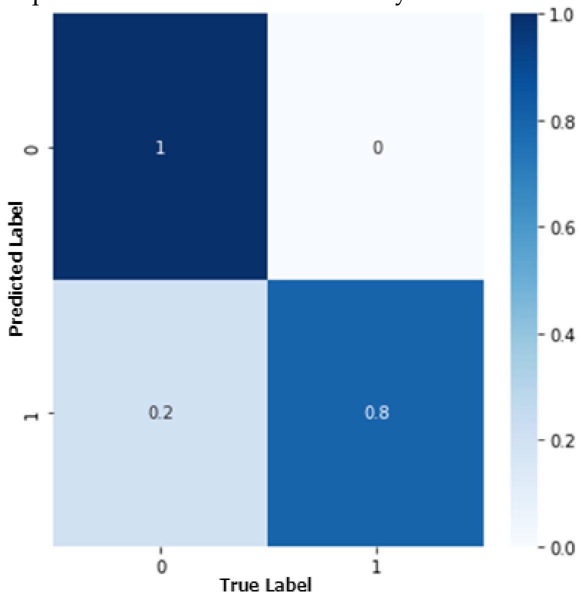

Figure 18 Confusion Matrix 0 represents suspicious movement, and 1 represents normal movement.

4. Discussion

The study was prepared with the aim of studying the suspicious movement that precedes the occurrence of theft in stores, and therefore the system studies two cases, the first is the suspicious movement, and the second is a normal movement, depending on the methodology that was followed during the development phase of the system and the pretreatment and division of the videos, the proposed system is generalizable And the study of individual suspicious movement or group suspicious movement (collective planning to carry out the theft).

With the system tested on many external videos that are not included in the dataset, the system was able to study, analyze and detect the individual suspicious movement of one person successfully, as well as in a place that includes many people and has a unique suspicious movement of a specific person, the system was able to identify, analyze and detect that movement, as well as For two people planning to carry out the theft, whether they are in the monitored area alone, or with a group of other people whose movement is normal.

The proposed model provided a high possibility of detecting prior planning to carry out theft operations (detection of suspicious movement).

We can expand the system in the future, by studying the intelligence of the group and detecting people who seek to plan a theft by identifying these people, which helps the security services in increasing the accuracy of tracking and monitoring.

The programming language Python and the TensorFlow software package were used to build the proposed neural network, in addition to using Opencv to process visual clips before entering them into the proposed neural network.

Regarding processing time, we use Google Colaboratory for experiments in this work. This tool is based on Jupyter Notebooks hardware and allows the use of a graphics processing unit (GPU). The training speed depends on the learning rate used, as it took 4 hours to train a neural network with a learning rate of 0.00001.

5. Conclusions

The current research presented a study of the structure of a neural network capable of analyzing and detecting suspicious movement (the movement that precedes the theft process), by using the dataset that we have, in addition to all the many videos on YouTube, we were able to propose a model capable of studying the sequence of several frames (160 frames) to identify and detect the nature of the movement within that period that represents (160 frames), the proposed model can provide great assistance to many stores in facilitating follow-up and monitoring, through the alerts that the model will appear when identifying a suspicious movement, and can also be used The system aims to provide the ability to review many of the previous video recordings to detect thefts that were not identified in the past.

References

1. Federation, N.R. 2020 National Retail Security Survey; National Retail Federation: Washington, DC, USA, 2020.
2. Guillermo A. Martínez-Mascorro, J. R.-P.-B.-C.-M. (2021, 9, 24). Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks. MDPI.
3. Carlos Ismael Orozco, M. E. (2020). CNN–LSTM Architecture for Action Recognition in Videos.
4. Noor Almaadeed, O. E.-M. (15 Mar 2021). A Novel Approach for Robust Multi Human Action Recognition and Summarization based on 3D Convolutional Neural Networks.
5. Wei Xu, M. Y. (n.d.). (2021).3D Convolutional Neural Networks for Human Action Recognition. USA.
6. Meriem Zerkouk, B. C.(2020). Spatio-Temporal Abnormal Behavior Prediction in Elderly Persons Using Deep Learning Models. Canada: MDPI.
7. S. Oprea, P. M.-G.-G.-V.-E.-R. (15 Apr 2020). A Review on Deep Learning Techniques for Video Prediction.
8. Park, Y.-H. K.-G. (2020). Predicting Future Frames using Retrospective Cycle GAN. IEEE XPLORE.
9. University of Central Florida. UCF-Crime Dataset. 2018. Available online: https://webpages.uncc.edu/cchen62/dataset.html (accessed on 23 April 2019).
10. Jeff Donahue, L. A. (31 May 201). *Long-term Recurrent Convolutional Networks for Visual Recognition and Description.*

11. J, R. T. (2020, Sep 2). *LSTMs Explained: A Complete, Technically Accurate, Conceptual Guide with Keras.* Retrieved from medium: https://medium.com/analytics-vidhya/lstms-explained-a-complete-technically-accurate-conceptual-guide-with-keras-2a650327e8f2

12.