IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 10, 2022, accepted May 23, 2022, date of publication May 27, 2022, date of current version June 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3178609

Spatiotemporal Activity Semantics Understanding Based on Foreground Object Segmentation: iCounter Scenario

TZU-WEI YU[1], MUHAMMAD ATIF SARWAR[1], YOUSEF-AWWAD DARAGHMI[2], SHENG-HSIEN CHENG[1], TSÌ-UÍ İK[1], (Member, IEEE), AND YIH-LANG LI[1], (Member, IEEE)

[1]Department of Computer Science, College of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan
[2]Department of Computer Systems Engineering, Palestine Technical University—Kadoorie, Tulkarem P310, Palestine

Corresponding author: Tsì-Uí İk (cwyi@nctu.edu.tw)

This work of T.-U. İk was supported in part by the Ministry of Science and Technology, Taiwan under grants MOST 110-2627-H-A49 -001, MOST 110-2221-E-A49-063-MY3, and MOST 111-2622-E-A49-009.This work was financially supported by the Center for Open Intelligent Connectivity from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

ABSTRACT Foreground object segmentation that captures the spatial and temporal information of moving objects in video is the most fundamental task for activity understanding in many intelligent applications, such as smart stores. Recently, several methods are proposed for the detection and recognition of activity based on object segmentation. However, these methods are often inaccurate because they do not maintain the temporal associations of object segment consistency across time. In this work, we proposed a hierarchical approach for foreground object segmentation and activity semantics understanding from sequential video to preserve spatial and temporal connectivity in the frames. The proposed system consists of two main modules: (a) the concatenated deep learning network containing PSPNet and convolutional-GRU to segment the foreground of an object of interest; (b) the activity mining framework which incorporates three sub-modules (i) a RetinaNet-based frame classifier to detect and count objects of interest; (ii) a time-domain activity and event detection algorithm; (iii) an image-based item query engine to recognize the shopping items. To evaluate the proposed approach, we designed the smart checkout-box called iCounter to collect the shopping activities dataset named ''NOL-41'' which is used in extensive experiments. The results show that the accuracy of the foreground object segmentation is 90.6%, the accuracy of the frame classification is 93.4%, the accuracy of activity event detection is 98.4%, and the accuracy of item query is 94.3%. Finally, the overall accuracy of the shopping list is 95.2%.

INDEX TERMS Activity semantics, activity recognition, self-checkout, PSPNet, Conv-GRU, image query, smart store, RetinaNet, image database, foreground object segmentation, video foreground segmentation.

I. INTRODUCTION

Foreground object segmentation is an active area of research in both computer vision and pattern recognition and has applications in different domains, such as autonomous driving [1], vehicle tracking [2], video surveillance [3], [4], real-time tracking [5], healthcare systems [6], internet of things (IoT) smart stores and content-based image retrieval [7]–[12]. Due to its ability to capture the spatial and temporal information of moving objects in a video, it can be used for activity understanding [13], [14]. Smart stores, which have become an important business recently, benefit largely from foreground object segmentation. Smart stores require capturing the temporal and spatial information of in-store activities to track product flow from shelves, to customers, and cashiers. Moreover, smart stores have a context that requires handling large inputs for real-time recognition of products and activities and real-time response to avoid any misdetection of segmentation or inconvenience in spatial and temporal connectivity.

Human activity understanding methods (Neural networks, Machine Learning Models) based on traditional handcrafted features depend on prior knowledge and human ingenuity to extract discriminating features [15], [16]. These methods can be broken down into three majors steps: (1) Foreground object detection that corresponds to activity segmentation, (2) Feature extraction and selection by domain knowledge

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

or an expert, and (3) Detection and recognition of activity understanding [17]–[19]. However, these methods (e.g. [20], [21]) are not sufficient for smart store spatiotemporal activity understanding based on the foreground object segmentation, particularly, when different objects share pixel region in the time domain in a sequential video. For example, when a shopper placed the item with their hand in the checkout box, the item's shared some region with their hand. When foreground object segmentation is applied to these objects, it doesn't properly work on it which leads to the misdetection of items in the real-time smart store.

These challenges (shared pixel) have been explored by many classical and modern approaches [22]–[24]. The optical flow was used in previous methods to maintain the temporal associations of object segments for the pixel consistency across the time for smoothness, but its temporal associations are inaccurate [25]. Long *et al.* in [26] proposed a fully convolutional network that yields a coarse heat-map for improving the only single image segmentation, but it does not apply to video segmentation. Pavel *et al.* [27] and Siam *et al.* [28] proposed a network for video segmentation that uses a combination of FCN and recurrent neural networks. Apart from these methods, there is a need for a method that covers foreground segmentation frame-by-frame with maintaining the spatiotemporal information consistent for large inputs in real-time [29], [30]. More recently, deep learning methods have attracted a lot of interest from the computer vision community. Because they can automatically learn representations from raw data and preserver the spatial and temporal information of objects in a sequential manner.

In this work, we propose a novel hierarchical approach for real-time foreground object segmentation and activity semantics understanding from sequential video to preserve spatial and temporal shared pixel connectivity "such as item in hand" in the frames [29]. The proposed approach consists of two main modules: (a) a concatenated deep learning network based on Pyramid Scene Parsing Network (PSP-Net) [30] to segment the foreground objects of interest, and convolutional-GRU [28] to preserve spatial and temporal connectivity of foreground objects in the frames; (b) the activity mining framework which includes three sub-modules (i) a RetinaNet-base [31] frame classifier to detect and count objects of interest; (ii) a time-domain activity and event detection algorithm; (iii) an image-based item query engine to recognize the shopping items. We also propose a design of a self-checkout box called iCounter with a camera that is mounted at a fixed angle to capture the spatiotemporal shopping activities information of the shopper as shown in figure 1. The iCounter is used to analyze the proposed hierarchical approach. We also contribute an NoL-41 video dataset of shopping objects. The dataset contains 25 videos of checkout activities. Videos are named V01∼V25 and captured by a Logitech C930e camera with a resolution of 640 × 480 and a frame rate of 15 fps. The total length of the videos is 534 seconds.

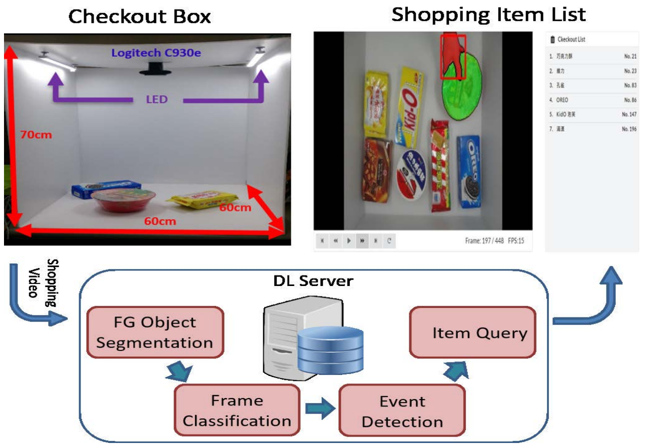


FIGURE 1. iCounter system architecture framework.

The proposed approach including the moving object segmentation and activity mining framework is evaluated by the intersection-over-union (IoU) and accuracy metric of each module in a real-time retail application scenario. The performance evaluation of the foreground object segmentation network was evaluated by three-fold cross-validation. There are 4216 frames including different glove colors, light intensity levels, and background patterns used for segmentation network training and evaluation. The foreground object segmentation module achieves an accuracy of 90.6%. The RetinaNet-based frame classification module has an accuracy of 93.4%. The accuracy of the shopping event module is 98.4% without considering the commodity category, the average error of the checkout event time is 0.018 seconds. The accuracy of the image query module is 95.2% considering the commodity category.

Our primary contributions and novelty of the proposed framework are: (i) Proposing an activity semantics understanding framework based on the foreground object segmentation and Conv-GRU to preserve spatial and temporal connectivity in the frames to analyze the shopping data. (ii) Proposing an activity mining framework that includes three sub-modules to detect, count, and recognize the activity's objects of interest in the time domain to smooth the shopping checkout process that reduces the timing of the retailer checkout process. (iii) Creating an NoL-41 video dataset of checkout shopping activity that will help the computer vision community to do further research on the checkout process. The claimed contributions address the shared pixel challenges in the smart store especially when the shopper adds the item into the iCounter for self-checkout. The foreground network keeps the information of each hand and item until the item is added to the shopping list. This is also beneficial for retail stores to avoid any misdetection and protect the retailers' investment from any big loss in the future. The rest of the paper is organized as follows: Section II describes the related work about segmentation, object detection, and recognition. Section III presents the methodology and prototype design. Section IV describes the performance evaluation and implementation environment, and section V concludes this research.

II. RELATED WORK

Retail organizations have a major concern about the item's billing in retailer stores. If customers go for the checkout process and find misdetection, missing item, or extra item in their total bill, they. will be distracted from the store which worsens the store's reputation. This causes hurt the progress and make a reason of big loss for retailer stores [32]. To avoid such incidents, different technologies were proposed including machine learning, deep learning, and computer vision technologies. Moreever, these papers [33]–[35] explore the salient object detection in the field of the internet of things which help to deploy the smart store solution.

A. CLASSICAL MACHINE LEARNING APPROACHES

Traditional foreground object segmentation approaches often use functions such as Gaussian Mixture Model (GMM) [36] and Gaussian distribution [37], [38]. The Gaussian distribution is used to calculate the background pixels, which is specifically used for video analysis, that is, learning the environment of each frame and comparing different frames, and storing the previous frames. This time-lapse method improves the results of motion analysis. However, noise is often generated due to changes in light, and the performance parameters also affect the result of the object segmentation.

B. DEEP LEARNING APPROACHES

In recent years, the development of deep learning has achieved great success in semantic segmentation. The initial deep learning method applied to image segmentation is patch classification. The patch classification method slices images into pieces for feeding to the depth model because the fully connected layer requires a fixed-size image. A fully convolutional network (FCN) replaced the full connection layer with convolution [39], [40], so the input size can be variable, and the speed is faster. However, there is still a problem with semantic segmentation, which is the down-sampling operations [41].

The down-sampling pooling operation was solved by two different models. Firstly, the FCN-based encoder-decoder architecture reduces the spatial dimension due to pooling, and the decoder gradually restores the spatial dimensions. There are usually cross-layer connections from an encoder to a decoder. The networks belonging to the encoder-decoder architecture are U-Net [42] and SegNet [43]. The second is dilated convolution architecture that replaces the pooling and maintains spatial resolution [44]. It also integrates semantic information well because it can expand the receptive field. The DeepLab series [45]–[48] and Pyramid Scene Parsing Network (PSPNet) networks [30] belong to the dilated convolution. However, the aforementioned methods and technologies suffer from drawbacks such as noise, downsampling, and maintaining the time sequence pattern. These drawbacks reduce the accuracy, and therefore, there is a need for methods that analyze the streaming data and understand the activity semantics.

C. PROPOSED METHOD BACKGROUND

PSPNet is a foreground segmentation-based network that combines the concept of global average pooling and feature fusion to achieve semantic segmentation. The feature fusion is a pyramid structure, also known as Pyramid Pooling Module [49]. The Pyramid Pooling Module fuses features on four different pyramid scales [50]. Like PSPNet other foreground object segmentation can be used in different applications for analyzing videos and understanding activities in the video. Although smart stores require shopping activities understanding, the foreground object segmentation has not been fully utilized in this domain. For example, Amazon Go eliminates the distress of customer queues using self-checkout technology including deep learning and computer vision modules to analyze the shopping activities for the shopping process [51]. iStore smart store [7], [8] is based on computer vision and deep learning for smart shopping in which YOLOv2 is used for item recognition. Because the accuracy and performance of smart store systems can still be improved, so foreground segmentation technique (PSPNet) still needs an approach that maintains the frame-by-frame activity stream to enable the understanding of shopping activity semantics.

To preserve the spatiotemporal connective the GRU analysis the foreground object frame-by-frame. Gated Recurrent Unit (GRU) is an LSTM architecture that is commonly used to process time-series data and is suitable for the analysis of sequential frames. The GRU works on the same principle but with simpler architecture as compared to LSTM. The GRU uses two gates such as reset r_t and update z_t to capture the temporal relation of the signal within the cell. Equation 1 describes the mathematically model of the GRU where h is a hidden state, t is the current time step, x is the input, σ is the activation function and W is the weight.

$$r_t = \sigma(W_r x_t + W_r h_{t-1} + b_r) \tag{1a}$$
$$z_t = \sigma(W_z x_t + W_z h_{t-1} + b_z) \tag{1b}$$
$$\hat{h}_t = \tanh(W(r_t \cdot h_{t-1}) + x_t) \tag{1c}$$
$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \hat{h}_t \tag{1d}$$

In our previous papers, we examined the self-checkout process with the help of iCarts [7], [8], and iShelf [9]. iCart is a lightweight smart self-checkout solution, that utilizes a smartphone mounted on a shopping cart and connected to a back-end deep-leaning server. The deep learning network is used to analyze the context of each frame to classify and action recognition. iShelf is an on-shelf item tracking solution that combines the load cells with sensor fusion and deep learning techniques. The tracking process incorporates determining the position and the weight of an item on the shelf. We generated the datasets for both (iCart, iShelf) and tested them on the proposed approaches which achieved the state-of-the-art performance. Similarly, iCounter is also a part of the self-checkout series that uses the below-mentioned methodology.

III. THE PROPOSED APPROACH

This section presents the proposed hierarchical approach for real-time foreground object segmentation and activity semantics understanding. The section firstly shows the context, smart store, in which the proposed approach can be used and evaluated. Then, the section shows the composition of the proposed approach including the foreground object segmentation module and the activity mining framework.

A. iCounter SMART STORE PROTOTYPE

The prototype design of the iCounter system is shown in figure 1. As illustrated in the figure, a camera is mounted at top of the checkout box to capture the checkout activities of the shopper at the checkout counter. The intelligent engine is mainly divided into two modules which are explained step-by-step in Figure 2. Figure 2(a) shows the captured video of the checkout box at a 15 FPS. The symbol $s0 = (0, 0)$ indicates that no item and hand in the frame. Similarly, $s1 = (1, 0)$ indicates one hand and no item, and $s2 = (1, 1)$ indicates one hand and one item in the frame. Figure 2(b) shows the foreground object segmentation module. The main purpose of the network is to cut out the foreground object from the image and distinguish the hands and items in the foreground block.

Figure 2(c) to figure 2(e) illustrate the activity mining framework which include three sub-modules. Figure 2(c) is a frame classification that is mainly used to analyze the action state at each time instance and apply smoothing to correct some misjudged frame labeling. Figure 2(d) shows the action segmentation and event detection module. The detection is based on the change of the frame and the relative position of the hand detection. For example, when t = 4, an event of adding an item is detected because of the state transition of the frame. Figure 2(e), detects an event of adding an item at t = 4, and further queries the categories of items after the added items.

B. FOREGROUND OBJECT SEGMENTATION MODULE

Foreground object segmentation module has three parts: (i) semantic segmentation network that applies the segmentation on the foreground object in the sequential video, (ii) Convolutional gated recurrent unit that preserves the spatial connectivities of the frames in the sequential video, and (iii) the classifier that classifies the segmented objects including hand and item.

1) SEMANTIC SEGMENTATION NETWORK

The segmentation network is the first part of the foreground object segmentation module as shown in figure 3. The network uses the backbone of the pyramid partition network (PSPNet) to obtain the features of the semantic data from the input. The PSPNet combines the concept of global average pooling and feature fusion to achieve semantic segmentation. The feature fusion is a pyramid structure, also known as Pyramid Pooling Module. The segmentation network accepts the input $I_{w,h,d}$ by the original RGB image having 320 × 256 resolution as mentioned in equation 2. In equation 2, $I_{t-1(w,h,d)}$ defines the previous frame, $I_{t(w,h,d)}$ current frame, $I_{t+1(w,h,d)}$ next frame, where w, h, d and n describe the width, height, dimension and the total number of the frames, respectively.

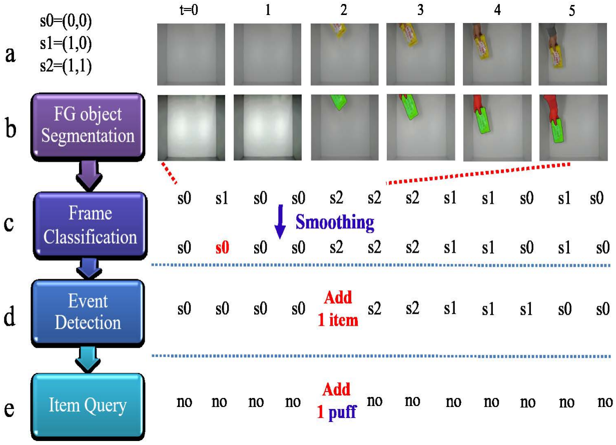

t=0
1
2
3
4
5
a
s0=(0,0)
s1=(1,0)
s2=(1,1)
b
FG object Segmentation
c
Frame Classification
s0 s1 s0 s0 s2 s2 s2 s1 s1 s0 s1 s0
Smoothing
s0 s0 s0 s0 s2 s2 s2 s1 s1 s0 s1 s0
d
Event Detection
s0 s0 s0 s0 Add 1 item s2 s2 s1 s1 s1 s0 s0
e
Item Query
no no no no Add 1 puff no no no no no no no

FIGURE 2. iCounter modules working flow.

$$I_{w,h,d} = \sum_{t=0}^{n} (I_{(t(w,h,d)-1)} + (I_{(t(w,h,d))} + (I_{(t(w,h,d)+1)}) \qquad (2)$$

The network generates the feature map as a output $x_{t(w,h,d)}$ in equation 3. The feature map length and width are 1/8 of the original image, and the depth is 128, which is the data block marked as a yellow square in figure 3. It is necessary to detect the foreground object through the information of the previous frame because the front and back scenes must be identified in time to find the changing area.

$$x_{t(w,h,d)} = PSPNet(I_{t(w,h,d)}) \qquad (3)$$

2) CONVOLUTIONAL GATED RECURRENT UNIT

The Convolutional Gated Recurrent Unit (Conv-GRU) is the second part of the foreground module which is used to process time-series data, and it is suitable for the analysis of continuous images. The Conv-GRU is embedded in the semantic segmentation network that has a combination with a convolutional network. The full architecture of the Conv-GRU network is shown in figure 4. The Conv-GRU is a pixel-level network. The Conv-GRU gets the inputs into two parts: The first part is the feature map output $x_{t(w,h,d)}$ of the PSPNet. The length and width of feature map $x_{t(w,h,d)}$ are 1/8 of the original image and depth 128 (yellow square in figure 4). The second is the historical feature map output of the previous image after passing through the Conv-GRU network. The length and width of the historical feature map are 1/8 of the original image and the historical feature map of depth is 64 (green square in figure 4). These two inputs are merged by concatenation and made two different 5 × 5 convolutions in sequence. Conv-GRU has a gating mechanism to regulate the flow of information like remembering the context of the previous and current frame. They use the reset and update gate to keep track of what information can be kept and what can be forgotten from the past. The reset gate (r) takes the input feature map $x_{t(w,h,d)}$ and previous feature map h_{t-1}, then applies the sigmoid activation function defined in equation 4a.

Similarly, the update gate (z) uses the same inputs and applies the sigmoid activation function defined in equation 4b. The reset gate stores the relevant content $h_t^{'}$ from the past which is calculated as in equation 4c. The reset gate and feature map h_{t-1} do the point-to-point multiplication and then the concatenation with the semantic feature map. After a 5 × 5 convolution, the model applies the *tanh* associated activation and outputs the current feature map h_t (blue square in figure 4) with a length and width are 1/8 of the original image and the depth is 64. The current feature map h_t is the product of z and memory content $\hat{h}_t$ plus the product of 1-z and the previous feature map h_{t-1}, which is also a previous feature map for the next frame as shown in equation 4d.

$$r_t = \sigma(W_r x_t + W_r h_{t-1} + b_r) \tag{4a}$$

$$z_t = \sigma(W_z x_t + W_z h_{t-1} + b_z) \tag{4b}$$

$$\hat{h}_t = \tanh(W(r_t \odot h_{t-1}) + x_t) \tag{4c}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \tag{4d}$$

3) CLASSIFIER

The classifier is the third part of the foreground module and uses the softmax activation. The classifier takes the current GRU feature map h_t as input, and generates a two-dimensional matrix of 320 × 256 as output, figure 3. The size of the matrix is equal to the original image size and the value of each position in the matrix is the predicted category for location. The classifier classifies the input feature map into three categories of foreground semantic objects as defined in equation 5.

$$Classifier = \begin{cases} background, & if\ y^t = 0 \\ hand\ (red), & if\ y^t = 1 \\ item\ (green), & if\ y^t = 2 \end{cases} \tag{5}$$

The proposed novel design for spatiotemporal activity semantics understanding based on foreground object segmentation allows the network to preserve the key information of each object even if the misdetection occurs during the segmentation process. For example, if the foreground network miss-detects either hand or item object in the frame sequence, the Conv-GRU will use the preserved hand and item information stored in the history unit (GRU) to compare the current and previous frames object information and does the appropriate tasks. The mechanism of the design network allows the activity mining framework to evaluate the modules including action detection and item recognition to maintain the state-of-the-art performance for iCounter shopping activity.

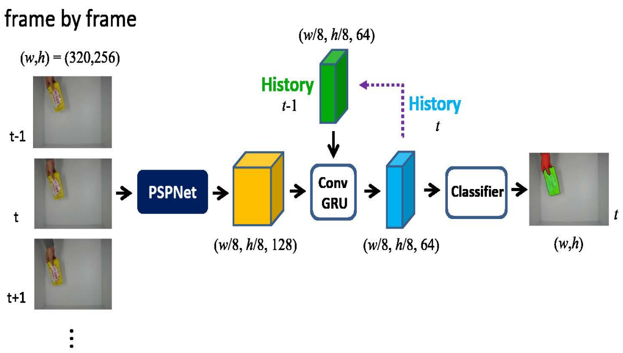


FIGURE 3. iCounter foreground object segmentation network.

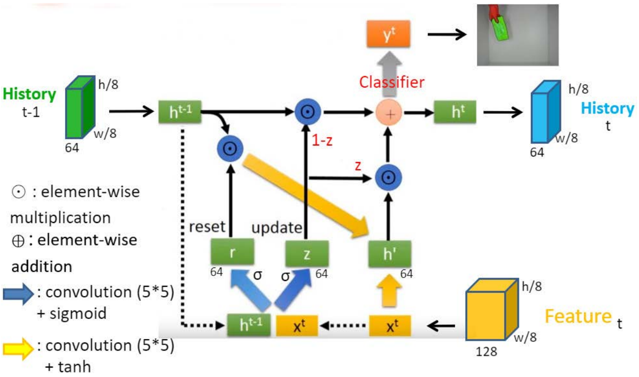


FIGURE 4. Convolution gated recurrent network for frame-to-frame process.

C. ACTIVITY MINING FRAMEWORK

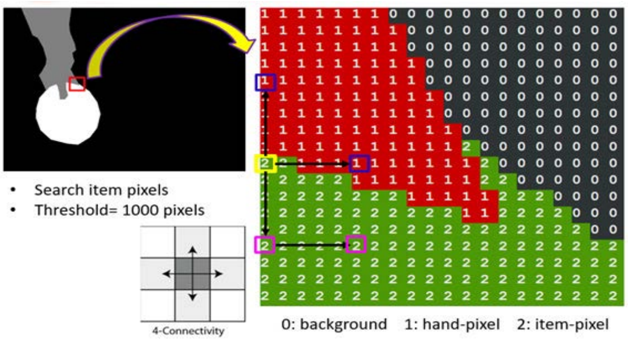
The activity mining framework consists of frame classification, action, and event detection, and item recognition modules. Frame classification module categories the frame w.r.t hands and items, the action and event detection module detects the events in the time domain and the item recognition module recognizes the item identity.

1) FRAME CLASSIFICATION

Frame classification module works based on section III-B. This module divides the frames into (m, n) categories based on the foreground object as defined in equation 6. Each frame has at most 2 hands (m) and each hand takes at most one hand-held item (n) per frame. Frame classification uses the RetinaNet network for hand bounding box detection and hand counting, and the item search algorithm finds the hand-held item in the frame.

$$Frame\ Categories = \begin{cases} 0 <= m <= 2, & m\ => \ hands \\ 0 <= n <= m, & n\ => \ items \end{cases} \tag{6}$$

The item search algorithm combines the concepts of depth-first search (DFS) and 4-connect for searching the hand-held items, as shown in Figure 5. As illustrated in the figure, the values 0, 1, and 2 represent background, hand, and hand-held items pixels, respectively. First, it searches the hand-held item



1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	2	0	0	0	0	0	0	0	0
2	2	1	1	1	1	1	1	1	1	1	1	2	0	0	0	0	0	0	0
2	2	2	2	2	1	1	1	1	1	1	1	2	2	0	0	0	0	0	0
2	2	2	2	2	2	2	2	1	1	1	1	1	2	2	2	0	0	0	0
2	2	2	2	2	2	2	2	2	2	2	1	1	2	2	2	2	0	0	0
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	0
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

FIGURE 5. Item location search algorithm working flow.

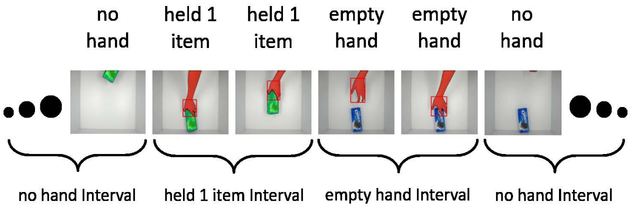


FIGURE 6. Action detection based on object segmentation.

pixel along with the hand mark bounding box. If the hand-held item pixel is searched, as shown in the yellow mark in the figure, then a depth-first search of 4 connections starts the search in 4 different directions such as top, bottom, left, and right. Depth-first Search each time jumps five pixels to improve the efficiency of the search. Finally, the left-most, right-most, top-most, and bottom-most positions are taken out from the pixels. If the area enclosed by the four boundaries is greater than the set threshold (1000 pixels), then the position of the item and the number of hand-held items are updated.

2) ACTION AND EVENT DETECTION

The action and event detection module is based on section III-C1. The frame classification contains the spatial information of the hand and the hand-held items which is helpful to detect the checkout event in the time domain. The event consists of two categories including added event and no event. Added event means that there is a hand to put an item into the checkout counter, and no event means there is an empty hand put into the checkout box. The event between two consecutive no-hand time intervals is judged by the action as shown in figure 6. Figure 7 left to right shows the events such as (0, 0) means there is no-event, (1, 1) means added 1 item in the checkout-box, and, $\Delta hand_t$ and $\Delta item_t$ show the change in number of hands and items. Similarly, (2, 2) means added 2 items. Action and event detection workflow is described in algorithm 1.

3) ITEM RECOGNITION

Item recognition [52] module recognizes the shopping item for a shopping list based on the event recognition. The item recognition network uses the foreground object image as input as shown in figure 8. Firstly, we create an image database containing various types of item images with different perspectives, occlusion, and reflection. The database has 400 item images. VGG16 deep learning model is used as feature extraction and gets the feature vector in the form of a histogram. The red histogram represents the item query image feature vector, and the blue represents the database feature vectors. Euclidean distance checks similarities between the database and query-image feature vector and obtained the final result.

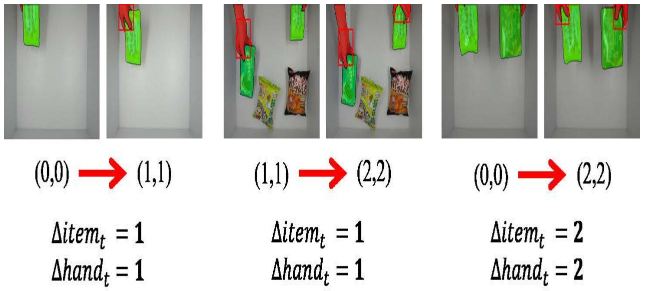


FIGURE 7. Activity event detection based on object segmentation.

Algorithm 1 Event Detection Algorithm

```
Require: Frame Classification Output
Ensure: Event Categories
  Initialization
  if $\Delta item_t > 0$ and $\Delta hand_t >= \Delta item_t$ then
      if $closestentrancehandwithitem$ then
          query images according to distance between hand and entrance
      end if
  end if
```

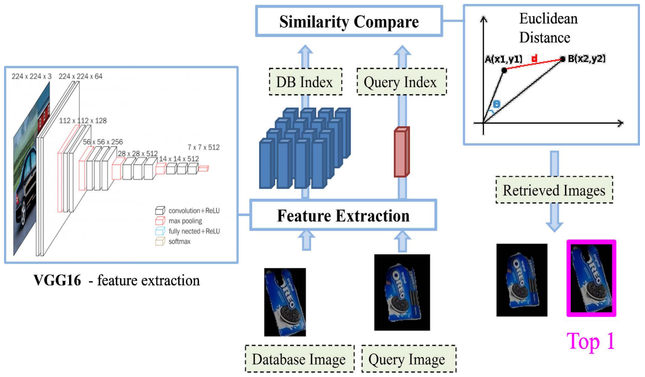


FIGURE 8. Image-based item query for item recognition.

IV. PERFORMANCE EVALUATION AND IMPLEMENTATION ENVIRONMENT

In this section, the proposed approach including the moving object segmentation and activity mining framework is evaluated. The moving object segmentation is evaluated by the

TABLE 1. Diversity of video V01∼V18.

Video	Glove Color	Light Intensity	Background Pattern
V01∼V06	NG	F	0
V07	B	F	0
V08	NG and B	F	1
V09	NG and B	F	2
V10	B	H	0
V11	B	H	1
V12	NG and B	H	2
V13	Bl and G	F	0
V14	Bl and G	F	1
V15	Bl and G	F	2
V16	Bl and G	H	0
V17	Bl and G	H	1
V18	Bl and G	H	2

intersection-over-union (IoU) metric and the activity mining framework is evaluated by the accuracy metric of each module.

A. NOL-41 DATASET

The dataset contains 25 videos of checkout activities. Videos are named V01∼V25 and captured by a Logitech C930e camera with a resolution of 640 × 480 and a frame rate of 15 fps. The total length of the videos is 534 seconds. At most two hands were involved in the checkout activities simultaneously, and at most one item was held by one hand. To increase the diversity of the dataset for model robustness and over-fitting prevention, videos were captured under various conditions, such as wearing gloves of different colors: black, blue, green, and no glove; different lighting levels: half lighting and full lighting; and different background patterns: white, single-color, and multi-colors. V01∼V18 are used for segmentation network training and performance evaluation, and V19∼V25 are used for the evaluation of activity recognition and shopping event detection.

Table 1 shows 18 videos with foreground labels that were used for segmentation network training and performance evaluation. In table 1, the "Glove Color" column represented the glove colors including black, blue, green, and no glove by B, Bl, G, and NG; the "Light Intensity" column represented the lighting levels, including half lighting and full lighting by H, F and "Background Pattern" column represented the background patterns including white, single-color, and multi-colors by 0, 1 and 2, respectively. Each video and its labeled data are stored in folders named video_01, video_02, . . . , etc. The hand and the hand-held item are foreground labeled objects in the image. The images are named 00000.jpg, 00001.jpg, etc. in the time domain and stored in the folder named Images, and foreground segmented objects images are stored in the Segment folder with the png extension. In addition, if there is no corresponding foreground object, it means just an empty background in the image.

The rest of the V19∼V25 videos consisting of 3384 frames in total were labeled with 198 checkout actions and 108 checkout events to evaluate the performance of the checkout event detection algorithm. Each video is labeled by the checkout action including action type, action start time, action end time, and action category. The action category depends on the hand action in the checkout box including empty-hand and hand-held items. Similarly, videos are labeled with checkout events including event start time, event category, and item name. The event has 2 categories including add and no event. Along with the labeling, the dataset also has a pixel distribution ratio for background, hand, and hand-held items for all frames. The background occupied a 92.3% area, while the area of hand and hand-held items are 4.9% and 2.8% respectively.

B. EVALUATION OF FOREGROUND OBJECT SEGMENTATION NETWORK

The performance evaluation of the foreground object segmentation network was evaluated by three-fold cross-validation. There are 4216 frames including different glove colors, light intensity levels, and background patterns used for segmentation network training and evaluation. The IoU metric is used to evaluate the network performance. To validate the over-fitting, we take glove color, light intensity, and background pattern as performance parameters for network evaluation.

The evaluation results of each parameter are shown in table 2. Table 2(a) shows that the model is sensitive to the color change of the glove especially when the model is evaluated using all glove colors images. So, more glove color data is needed for training and learning to improve the stability of the model. Table 2(b) shows that different background patterns haven't significantly changed the model evaluation. The black background has the highest accuracy rate for model evaluation. However, more data are needed for model training and evaluation, especially for multi-color backgrounds. Similarly, table 2(c) shows that half-light intensity has a more accurate result for model evaluation. The full light accuracy is mainly affected due to item shining packing and colors that cause the blurred image.

After parameters evaluation, the overall performance of the segmentation network was evaluated by three-fold cross-validation. Table 3 shows the overall accuracy of segmentation network is 90.6%. The average evaluation result shows that the segmentation network is the best fit for background and hand segmentation. However, more hand-held items labeled data are needed to increase average accuracy and boost model stability. Finally, we compared the proposed methodology with state-of-the-art (SOTA) foreground object segmentation methods as shown in table 4. Deeplab+GRU has closer accuracy to the proposed methodology but has slower fps. The proposed methodology has the highest accuracy for foreground object segmentation is 90.6% with a better fps rate.

C. EVALUATION OF FRAME CLASSIFICATION

Frame classification has a certain in-corrected frame label prediction that can be smoothed in the time domain before the shopping event detection. The smoothing effect is limited

TABLE 2. Evaluation of performance parameters.

(a) Glove Color Parameter Accuracy

Category	Background	Hand	Hand-held item	**Mean**
No-Glove	98.2%	85.4%	48.6%	**77.4%**
Black	98.8%	90.2%	71.3%	**86.8%**
Blue, Green	94.5%	15.3%	43.2%	**51.0%**
Average	**97.2%**	**63.6%**	**54.3%**	**71.7%**

(b) Background Pattern Parameter Accuracy

Category	Background	Hand	Hand-held item	**Mean**
blank	98.9%	92.3%	73.3%	**88.2%**
single-color	98.7%	90.6%	74.6%	**87.9%**
multi-color	98.7%	91.4%	69.2%	**86.5%**
Average	**98.8%**	**91.4%**	**72.4%**	**87.5%**

(c) Lighting Intensity Parameter Accuracy

Category	Background	Hand	Hand-held item	**Mean**
Full	98.7%	91.2%	70.1%	**86.7%**
Half	98.6%	90.3%	72.0%	**87.0%**
Average	**98.7%**	**90.8%**	**71.1%**	**86.9%**

TABLE 3. IoU evaluation of foreground object segmentation network (PSPNet+GRU).

Category	Background	Hand	Hand-held item	**Mean**
Folder 1	99.0%	93.1%	78.1%	**90.1%**
Folder 2	99.2%	92.6%	81.2%	**91.0%**
Folder 3	99.2%	92.5%	80.4%	**90.7%**
Average	**99.1%**	**92.7%**	**79.9%**	**90.6%**

TABLE 4. Comparative analysis with SOTA methods.

Method	Mean IOU	FPS
PSPNet	78%	21.2
PSPNet(3 frames)	78.2%	10.5
Segmenter [53]	81.3%	
SegFormer [54]	84.0%	2.5
Deeplab+GRU	87.4%	5.1
PSPNet+GRU	**90.6%**	10.1

to one frame before and after the reference frame. The major frame algorithm is used to smooth the current frame category. The major vote algorithm decides the current frame voting based on the previous and next frame category but only the current frame in the smoothing range.

Frame classification smoothing evaluation depends on the RetinaNet hand detection network and item search algorithm. RetinaNet was evaluated on 1168 foreground hand-labeled images that cheesed from V01∼V18 videos by three-fold cross-validation with the mAP metric. Table 5 shows the average accuracy of RetinaNet network is 97.1%. After evaluation, the major vote algorithm smooths the in-corrected frame label. Figure 9 shows the smoothing example of a frame label like how an algorithm smooths the in-corrected frame label. There are two rows of the frames from left to right named ''original'' and ''major vote''. The original frames row indicates the predicted frames labels result, and the major vote frames row shows the smoothed in-corrected frame labels. The accuracy of frame classification before smoothing is 0.931% and after smoothing is 0.932%.

TABLE 5. Evaluation of hand detection.

Dataset	mAP
Folder 1	94.5%
Folder 2	97.1%
Folder 3	97.7%
Average	**97.1%**

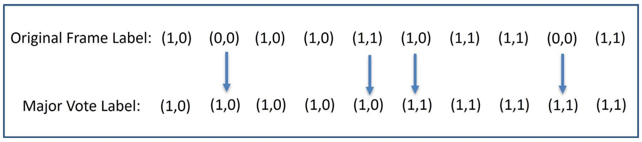

Original Frame Label: (1,0) (0,0) (1,0) (1,0) (1,1) (1,0) (1,1) (1,1) (0,0) (1,1)
Major Vote Label: (1,0) (1,0) (1,0) (1,0) (1,0) (1,1) (1,1) (1,1) (1,1) (1,1)

FIGURE 9. Smoothing algorithm for frame classification Mis-detection.

D. EVALUATION OF EVENT DETECTION AND ITEM RECOGNITION

The event detection module was evaluated on 3384 frames approximately that have 108 checkout events including 60 added and 48 no events by accuracy and time error metrics. The checkout events are detected by an algorithm 2. The event accuracy (EventAcc) is measured based on the pairing sequence between the ground truth and the predicted event in the time domain. The calculation formula of EventAcc is defined in equation 7. Table 6 shows that the event detection accuracy is 96.9% before smoothing and without recognizing the item category. The accuracy slightly increases after smoothing the event detection and accuracy is 98.4%.

The item query evaluation is based on the recognition of the corrected item category. The item image database has 32 item categories images with 53 additional items categories images for item query evaluation. The evaluation method compares the predicted item category with the actual category of items. The item query model selects the top k item images from the item image database. The top k represents the result of the item query image having the highest similarity with images of the item image database. If one of the top k results hits the real category, the item query image is considered correct. Table 7 shows the different top k item query results.

$$EventAcc = \frac{\text{\# of Correct Event Pairs}}{\text{\# of Correct Event Pairs} + \text{\# of unpaired events}} \tag{7}$$

E. IMPLEMENTATION ENVIRONMENT

The prototype system consists of a deep learning server and a checkout box. The main hardware architecture is shown in figure 1. The checkout box has a length and width of 60 cm and a height of 70 cm. A webcam (Logitech C930e)

Algorithm 2 Event Pairing Algorithm

```
Initialization
for $real\_event_i^t$ do
    for time = t-1 to t+1 do
        if $real\_event_i^t$ == $detect\_event_i^t$ then
            pair $real\_event_i^t$ with $detect\_event_i^t$
        end if
    end for
end for
```

TABLE 6. Evaluation of event detection.

Event Detection	Correct events	Incorrect events	Event Accuracy	Time Error
Before Smoothing	62	2	96.9%	0.018s
After Smoothing	62	1	98.4%	0.018s

TABLE 7. Evaluation of item query.

Item	Top 1	Top 5	Top 10
Correct	50	53	53
Incorrect	3	0	0
Incorrect	94.3%	100%	100%

is mounted on the top to capture the checkout activities video and uploaded them to the server. Two lamps are placed on the left and right sides of the checkout box to provide illumination. The deep learning server operating system is Ubuntu 18.04 LTS, and NVIDIA GEFORCE GTX 1080Ti GPU to accelerate the execution of deep learning algorithms, and finally, the results are displayed on the screen with web pages.

V. CONCLUSION

This paper has proposed a hierarchical approach for foreground object segmentation and activity semantics understanding from sequential video to preserve spatial and temporal connectivity in the frames. The proposed system handles large inputs for real-time recognition of products and activities and real-time response to avoid any misdetection of segmentation or inconvenience in the spatial and temporal domain. We conclude that spatiotemporal activity semantics understanding based on foreground object segmentation has a state-of-the-art accuracy to detect and recognize the retail product in real-time applications. Also, different modules are implemented in a proposed system to maintain the state-of-the-art accuracy for foreground objects in the spatiotemporal domain. Our future work will include testing the system in different store environments.

REFERENCES

[1] H. Naveed, F. Jafri, K. Javed, and H. A. Babri, "Driver activity recognition by learning spatiotemporal features of pose and human object interaction," *J. Vis. Commun. Image Represent.*, vol. 77, pp. 103–135, May 2021.

[2] W. S. K. Fernando, H. M. S. P. B. Herath, P. H. Perera, M. P. B. Ekanayake, G. M. R. I. Godaliyadda, and J. V. Wijayakulasooriya, "Object identification, enhancement and tracking under dynamic background conditions," in *Proc. 7th Int. Conf. Inf. Autom. Sustainability*, Dec. 2014, pp. 1–6.

[3] Y. Xu, D. Xu, S. Lin, T. X. Han, X. Cao, and X. Li, "Detection of sudden pedestrian crossings for driving assistance systems," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 3, pp. 729–739, Jun. 2012.

[4] C. I. Patel, S. Garg, T. Zaveri, and A. Banerjee, "Top-down and bottom-up cues based moving object detection for varied background video sequences," *Adv. Multimedia*, vol. 2014, pp. 1–20, Jan. 2014.

[5] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1577–1591, Jul. 2013.

[6] J. Liu, Y. Liu, Y. Cui, and Y. Q. Chen, "Real-time human detection and tracking in complex environments using single RGBD camera," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 3088–3092.

[7] H.-C. Chi, M. A. Sarwar, Y.-A. Daraghmi, K.-W. Lin, T.-U. Ik, and Y.-L. Li, "Smart self-checkout carts based on deep learning for shopping activity recognition," in *Proc. 21st Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Sep. 2020, pp. 185–190.

[8] M. A. Sarwar, Y.-A. Daraghmi, K.-W. Liu, H.-C. Chi, T.-U. Ik, and Y.-L. Li, "Smart shopping carts based on mobile computing and deep learning cloud services," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.

[9] M.-H. Lin, M. A. Sarwar, Y.-A. Daraghmi, and T.-U. Ik, "On-shelf load cell calibration for positioning and weighing assisted by activity detection: Smart store scenario," *IEEE Sensors J.*, vol. 22, no. 4, pp. 3455–3463, Feb. 2022.

[10] R. Trabelsi, I. Jabri, F. Smach, and A. Bouallegue, "Efficient and fast multimodal foreground-background segmentation using RGBD data," *Pattern Recognit. Lett.*, vol. 97, pp. 13–20, Oct. 2017.

[11] C. Cuevas, R. Martínez, and N. García, "Detection of stationary foreground objects: A survey," *Comput. Vis. Image Understand.*, vol. 152, pp. 41–57, Nov. 2016.

[12] F. J. López-Rubio and E. López-Rubio, "Features for stochastic approximation based foreground detection," *Comput. Vis. Image Understand.*, vol. 133, pp. 30–50, Apr. 2015.

[13] C. Li, Z. Li, Z. Ge, and M. Li, "Knowledge driven temporal activity localization," *J. Vis. Commun. Image Represent.*, vol. 64, Oct. 2019, Art. no. 102628.

[14] C. D. Nath and S. M. Hazarika, "Activity recognition in video sequences over qualitative abstracts of a diagram-based representation schema," *J. Vis. Commun. Image Represent.*, vol. 76, Apr. 2021, Art. no. 103061.

[15] D. R. Beddiar, A. Hadid, B. Nini, and M. Sabokrou, "Vision-based human activity recognition: A survey," *Multimedia Tools Appl.*, vol. 79, no. 41, pp. 30509–30555, 2020.

[16] A. Bux, *Vision-Based Human Action Recognition Using Machine Learning Techniques*. Lancaster, U.K.: Lancaster Univ., 2017.

[17] K. Appiah, A. Hunter, P. Dickinson, and H. Meng, "Accelerated hardware video object segmentation: From foreground detection to connected components labelling," *Comput. Vis. Image Understand.*, vol. 114, no. 11, pp. 1282–1291, 2010.

[18] M. Camplani and L. Salgado, "Background foreground segmentation with RGB-D kinect data: An efficient combination of classifiers," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 122–136, Jan. 2014.

[19] J. Gallego and M. Pardàs, "Region based foreground segmentation combining color and depth sensors via logarithmic opinion pool decision," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 184–194, Jan. 2014.

[20] A. Chriki, H. Touati, H. Snoussi, and F. Kamoun, "Deep learning and handcrafted features for one-class anomaly detection in UAV video," *Multimedia Tools Appl.*, vol. 80, no. 2, pp. 2599–2620, Jan. 2021.

[21] M. Vijayan and R. Mohan, "A universal foreground segmentation technique using deep-neural network," *Multimedia Tools Appl.*, vol. 79, nos. 47–48, pp. 34835–34850, Dec. 2020.

[22] L. Bao, B. Wu, and W. Liu, "CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5977–5986.

[23] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, ''SegFlow: Joint learning for video object segmentation and optical flow,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 686–695.

[24] P. Hu, G. Wang, X. Kong, J. Kuen, and Y.-P. Tan, ''Motion-guided cascaded refinement network for video object segmentation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1400–1409.

[25] P. Tokmakov, K. Alahari, and C. Schmid, ''Learning motion patterns in videos,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3386–3394.

[26] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[27] M. S. Pavel, H. Schulz, and S. Behnke, ''Recurrent convolutional neural networks for object-class segmentation of RGB-D video,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–8.

[28] M. Siam, S. Valipour, M. Jagersand, and N. Ray, ''Convolutional gated recurrent networks for video segmentation,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3090–3094.

[29] H. Wang, W. Liu, and W. Xing, ''Video object segmentation via random walks on two-frame graphs comprising superpixels,'' *J. Vis. Commun. Image Represent.*, vol. 80, Oct. 2021, Art. no. 103293.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, ''Pyramid scene parsing network,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, ''Focal loss for dense object detection,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[32] G. D. Jyothi and K. Navya, ''Design and implementation of a store management system,'' in *Proc. Int. Conf. Intell. Sustain. Syst. (ICISS)*, Dec. 2017, pp. 1149–1151.

[33] Z. Gao, H. Zhang, S. Dong, S. Sun, X. Wang, G. Yang, W. Wu, S. Li, and V. H. C. de Albuquerque, ''Salient object detection in the distributed cloud-edge intelligent network,'' *IEEE Netw.*, vol. 34, no. 2, pp. 216–224, Mar. 2020.

[34] Z. Gao, C. Xu, H. Zhang, S. Li, and V. H. C. de Albuquerque, ''Trustful internet of surveillance things based on deeply represented visual co-saliency detection,'' *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4092–4100, May 2020.

[35] J. Zhang, C. Xu, Z. Gao, J. J. P. C. Rodrigues, and V. H. C. de Albuquerque, ''Industrial pervasive edge computing-based intelligence IoT for surveillance saliency detection,'' *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 5012–5020, Jul. 2021.

[36] P. KaewTraKulPong and R. Bowden, ''An improved adaptive background mixture model for real-time tracking with shadow detection,'' in *Video-Based Surveillance Systems*. Boston, MA, USA: Springer, 2002, pp. 135–144.

[37] A. B. Godbehere, A. Matsukawa, and K. Goldberg, ''Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation,'' in *Proc. Amer. Control Conf. (ACC)*, Jun. 2012, pp. 4305–4312.

[38] Z. Zivkovic, ''Improved adaptive Gaussian mixture model for background subtraction,'' in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 2, 2004, pp. 28–31.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ''ImageNet classification with deep convolutional neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[40] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, ''A benchmark dataset and evaluation methodology for video object segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 724–732.

[41] H. Wu and X. Gu, ''Max-pooling dropout for regularization of convolutional neural networks,'' in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2015, pp. 46–54.

[42] O. Ronneberger, P. Fischer, and T. Brox, ''U-net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[43] V. Badrinarayanan, A. Kendall, and R. Cipolla, ''SegNet: A deep convolutional encoder-decoder architecture for image segmentation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[44] F. Yu and V. Koltun, ''Multi-scale context aggregation by dilated convolutions,'' 2015, *arXiv:1511.07122*.

[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, ''Semantic image segmentation with deep convolutional nets and fully connected CRFs,'' 2014, *arXiv:1412.7062*.

[46] L. C. Chen, G. Papandreou, and I. Kokkinos, ''DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2016.

[47] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, ''Rethinking atrous convolution for semantic image segmentation,'' 2017, *arXiv:1706.05587*.

[48] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, ''Encoder-decoder with atrous separable convolution for semantic image segmentation,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[49] K. He, X. Zhang, S. Ren, and J. Sun, ''Spatial pyramid pooling in deep convolutional networks for visual recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[50] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[51] Amazon. (2016). *Amazon Go*. Accessed: Jan. 23, 2017. [Online]. Available: https://www.amazon.com/b?ie=UTF8&node=16008589011

[52] *Image Retrieval*. Accessed: Dec. 26, 2018. [Online]. Available: https://github.com/willard-yuan/flask-keras-cnn-image-retrieval

[53] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, ''Segmenter: Transformer for semantic segmentation,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.

[54] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, ''SegFormer: Simple and efficient design for semantic segmentation with transformers,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–14.

TZU-WEI YU is currently pursuing the master's degree with the National Yang Mining Chiao Tung University, Taiwan. His research interests include the artificial intelligence, deep learning, and computer vision.

MUHAMMAD ATIF SARWAR received the B.S. and M.S. degrees in computer science from COMSATS University Islamabad, Sahiwal Campus, Pakistan, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree with the EECS International Graduate Program, National Yang Ming Chiao Tung University, Taiwan. His research interests include artificial intelligence, deep learning, and computer vision. His current research to detect activity recognition and actions in a retailers store, sports, and exercise.

YOUSEF-AWWAD DARAGHMI received the B.E. degree in electrical and computer engineering from An-Najah National University, in 2002, and the master's and Ph.D. degrees in computer science and engineering from the National Chiao Tung University, Taiwan, in 2007 and 2014, respectively. He is currently an Associate Professor at the Computer Systems Engineering Department, Palestine Technical University—Kadoorie. His research interests include intelligent transportation systems, vehicular *ad-hoc* networks, and blockchain. He received the Best Paper Award from the International Conference on Intelligent Transportation Systems Telecommunications, in 2012. He served as a Technical Program Committee Member for the International Conference on Connected Vehicles and Expo (ICCVE 2012–2016), the International Conference on Intelligent Transportation Systems Telecommunications (ITST 2012–2018), the International Conference on Signal Processing (ICOSP 2015 and 2016), and the Asia–Pacific Network Operation and Management Symposium (APNOMS 2015 and 2016). He is a Reviewer for some highly distinguished journals including IEEE Transactions on Intelligent Transportation Systems, *IEEE Communication Magazine*, and *IEEE Network* magazine.

SHENG-HSIEN CHENG is currently pursuing the master's degree with National Yang Mining Chiao Tung University, Taiwan. His research interests include artificial intelligence, deep learning, and computer vision.

TSÌ-UÍ İK (Member, IEEE) received the B.S. degree in mathematics and M.S. degrees in computer science and information engineering from National Taiwan University, in 1991 and 1993, respectively, and the Ph.D. degree in computer science from the Illinois Institute of Technology, in 2005.

He is currently a Professor with the Department of Computer Science and the Director of the Institute of Computer Science and Engineering, National Yang Ming Chiao Tung University. His research interests include intelligent applications, such as intelligent sports learning and intelligent transportation systems, mobile sensing, machine learning, deep learning, and wireless sensor and *ad-hoc* networks.

Dr. İk had been a Senior Research Fellow of the Department of Computer Science, City University of Hong Kong. He was bestowed the Outstanding Young Engineer Award by the Chinese Institute of Engineers, in 2009, and the Young Scholar Best Paper Award by IEEE IT/COMSOC Taipei/Tainan Chapter, in 2010. He received the Best Paper Award at ITST 2012. He received a three-year Outstanding Young Researcher Grant from the National Science Council, Taiwan, in 2012. In 2020, he received the Sports Science Research and Development Award, MoE, Taiwan. In 2020 and 2021, his research works received the MOST Future Tech Award.

YIH-LANG LI (Member, IEEE) received the B.S. degree in nuclear engineering and the M.S. and Ph.D. degrees in computer science (majoring in designing and implementing a highly parallel cellular automata machine for fault simulation) from the National Tsing Hua University, Hsinchu, Taiwan. From 1995 to 1996 and from 1998 to 2003, he was a Software Engineer and an Associate Manager with Springsoft Corporation, Hsinchu, where he first completed the development of design rule checking (DRC) tool for the custom-based layout design and then established and led a routing team for developing a block-level shape-based router for the custom-based layout design. In 2003, he joined the Faculty of the Department of Computer Science, National Yang Ming Chiao Tung University (NCTU), Hsinchu, where he is currently a Professor. His current research interests include physical synthesis, parallel architecture, vehicle navigation, and deep learning. He joined the Technical Committee of the first CAD contest in Taiwan and served as a Committee Member for ten years. He has also been serving as the Compensation Committee Member and the Independent Director of Board of Directors for AMICCOM Electronics Corporation, since 2012. He was a recipient of the Japan Society for the Promotion of Science Faculty Invitation Fellowship. He was the Contest Chair of the first CAD contest at ICCAD, in 2012, and the Technical Program Committee Member for ASPDAC and DAC.