SMRT: Surveillance Monitoring and Recognition Techniques for Analyzing Service Behavior in Blurred and Unsteady Video

Qiaoyun Zhang[1], Shih-Yang Yang[3], Yu Lin[4], Huan-Chao Keh[2], Diptendu Sinha Roy[5]*

[1] School of Artificial Intelligence, Chuzhou University, China
[2] Department of Computer Science and Information Engineering, Tamkang University, Taiwan
[3] Department of Media Arts, Kang-Ning University, Taiwan
[4] Department of Nursing, Kang-Ning University, Taiwan
[5] Department of Computer Science and Engineering, National Institute of Technology Meghalaya, India
zqyun@chzu.edu.cn, Taiwan.shihyang@g.ukn.edu.tw, yulin@ukn.edu.tw, hckeh@mail.tku.edu.tw, diptendu.sr@nitm.ac.in

Abstract

With the rapid development of the service industry and increasing customer expectations, traditional mystery shopper audit methods face several challenges, such as time-consuming manual analysis, significant subjective bias, and difficulty in accurately quantifying complex service behaviors. To overcome these limitations, this paper introduces an innovative approach called Surveillance Monitoring and Recognition Techniques (SMRT) for analyzing service behavior. The proposed SMRT achieves precise classification of service behaviors through a two-phase process: coarse-grained and fine-grained analysis. In the coarse-grained phase, the proposed SMRT preprocesses blurred video to extract and emphasize relevant external features, specifically detecting and capturing 'person' objects in video frames, thereby effectively filtering out irrelevant frames and reducing computational load. In the fine-grained phase, it performs spatiotemporal feature extraction and utilizes Transformer models to conduct a detailed comparison of target behavioral features across video segments. Simulation results demonstrate that the proposed SMRT significantly enhances recognition performance in terms of accuracy, and F1-score compared to existing methods.

Keywords: Blurred and unsteady video, Mystery shopper audit, Service behavior recognition

1 Introduction

In an era of increasingly fierce business competition, companies must continuously improve and innovate to meet customer demands and maintain a competitive edge. To ensure the highest standards of customer service, a mysterious role has quietly emerged in the retail and service industries, known as 'mystery shoppers.' These individuals discreetly evaluate customer service interactions, providing companies with critical insights into the quality of their services. However, the videos recorded during these evaluations often present significant challenges for automated analysis, such as camera shake, blurriness, complex backgrounds, and low resolution.

With the rapid development of Artificial Intelligence (AI) technology, the application of image processing techniques in various industries is increasingly expanding [1-3]. Technologies such as Faster Region-Convolutional Neural Networks (Faster R-CNN) [4], You Only Look Once (YOLO) series [5-7], Single Shot Multibox Detector (SSD) [8], and others demonstrate fast and high accuracy in object detection, making them commonly used and compared in various tasks. In behavior recognition, the usual objective is to identify and understand actions and patterns of motion in videos and to classify or score the accuracy of these actions. Technologies such as two-stream CNN [11-14], 3DCNN [15-17], and Long Short-Term Memory (LSTM) [19-20], among others, are utilized for this purpose. However, achieving both fast and high-precision results requires high-quality datasets, which are crucial for the effectiveness of these techniques.

To address these issues, this paper introduces a behavior recognition mechanism called SMRT. The proposed SMRT leverages the capabilities of multi-model to identify the behavior of service providers in mystery shopping audit videos. Initially, it preprocesses blurred video to extract and emphasize relevant external features for 'person' object detection to filter out irrelevant frames. It then extracts key behavioral features, followed by the Transformer model [10] to analyze temporal features. The contributions of this paper are detailed as follows:

1) Effectively Filtering out Irrelevant Frames

The proposed SMRT utilizes the coarse-grained phase to filter out irrelevant frames, reducing computational load and allowing the system to focus on frames that contain actual service behaviors. This strategy significantly enhances processing efficiency, particularly in long or low-frame-rate videos, and addresses video blurring issues.

2) Employing a Transformer to Capture Temporal Dependencies

The proposed SMRT utilizes the Transformer model to analyze the temporal dependencies between video frames in fixed-duration segments. This allows for the identification of complex behavioral patterns and accurate

*Corresponding Author: Huan-Chao Keh; E-mail: hckeh@mail.tku.edu.tw
DOI: https://doi.org/10.70003/160792642025052603007

classification of service behaviors, providing detailed assessments of service quality.

3) Integrating Multimodal Feature Fusion

The proposed SMRT emphasizes the fusion of visual and temporal features, leveraging both static and dynamic information for behavior recognition. By integrating multiple models, SMRT captures both the static features of individual frames and the temporal variations across frame sequences, enabling a comprehensive and precise analysis of service behaviors.

The remainder of the paper is structured as follows: Section 2 reviews related work. Section 3 outlines the problem and research objective. Section 4 details the proposed SMRT. Section 5 evaluates its performance and Section 6 summarizes the key findings and future research directions.

2 Related Work

This section reviews existing research relevant to the proposed SMRT mechanism, categorized into two areas: behavioral identification based on image analysis and skeleton data.

2.1 Behavioral Identification Based on Image Analysis

Deep learning methods like two-stream CNN [11-14], 3DCNN [15-17], and LSTM [19-20] have been widely used for behavior identification. Simonyan and Zisserman [11] introduced a two-stream CNN for separating spatial and temporal features in videos. Khan et al. [12] improved spatial stream robustness through dataset augmentation techniques like rotation and flipping to mitigate overfitting. Wang et al. [13] integrated bidirectional gated recurrent unit into the two-stream CNN, capturing temporal dependencies across frames for nuanced action recognition. However, these approaches are incapable of distinguishing some roughly similar actions in videos.

To address this, Zhou et al. [14] introduced a Multi-head Attention-based Two-stream EfficientNet, combining multi-head attention and two-stream EfficientNet for key action recognition. While effective for short-term dependencies, it faced limitations in handling longer video sequences. Tran et al. [15] introduced 3DCNN for spatiotemporal feature learning but suffered from high memory consumption due to treating spatial and temporal dimensions equally. Zhou et al. [16] and channel attention-based 3DCNN [17] reduced computational overhead and enhanced feature selection, but frame-by-frame processing remained inefficient.

Tran et al. [15] proposed a method for spatiotemporal feature learning using 3DCNN trained on a large-scale supervised video dataset. However, they treated the temporal and spatial dimensions equally, leading to significant memory consumption in practical applications. To address this issue, Zhou et al. [16] and Zhao et al. [17] employed 3DCNN to reduce computational overhead and enhance feature selection, but frame-by-frame processing remained inefficient.

LSTM-based approaches [18-20] focused on temporal feature modeling. Ng et al. [18] utilized a dual-stream network with LSTM to extract spatial and optical flow features, while Li et al. [19] incorporated attention mechanisms into LSTM to leverage spatial correlations. Although effective in improving feature representation, these methods struggled with computational efficiency for long-duration videos. Dai et al. [20] employed two-stream attention-based LSTM but faced similar efficiency challenges.

To address these limitations, the proposed SMRT introduces a pre-filtering step by detecting 'person' objects, focusing on frames with relevant behaviors and reducing computational overhead. By extracting key behavioral features and employing a Transformer model for temporal analysis, it captures dependencies across frames, enabling accurate recognition of complex and extended behaviors.

2.2 Behavioral Identification Based on Skeleton Data

Skeleton data, compared to RGB and depth data, is less influenced by background noise, lighting, and appearance variations, making it effective for behavior recognition. Yan et al. [21] introduced Skeleton-based Spatiotemporal Graph Convolutional Network, utilizing graph convolutional networks to capture spatiotemporal relationships in skeleton sequences. However, its local convolution structure failed to fully exploit global joint relationships, limiting its ability to model complex, long-duration actions involving co-movement of distant joints.

To enhance feature representation, Wu et al. [22] proposed the Multi-grain Contextual Focus module, which captured relational information between body parts and joints, providing more interpretable skeleton representations. Yin et al. [23] introduced a lightweight double-feature triple-scale motion network to improve efficiency and accuracy. However, in real-world scenarios, videos often suffer from poor image quality, leading to inaccuracies in skeleton detection and recognition.

The proposed SMRT addresses these challenges by integrating visual features from RGB frames with temporal dependencies. This approach combines the strengths of image-based and skeleton-based methods, providing a comprehensive analysis of service behaviors, even in complex and noisy environments.

3 Notations, Assumptions and Problem Descriptions

This section introduces the notations, assumptions, problem descriptions, and objective.

3.1 Notations and Assumptions

Mystery shopping is a popular strategy for businesses seeking to gain a competitive edge. Mystery shoppers, acting as regular consumers, discreetly assess products and services, often using concealed cameras to provide authentic feedback.

A mystery shopping audit video V consists of m continuous segments $V = (v_1, v_2, \ldots, v_m)$, where each $v_i \in V$ is a fixed-duration segment. The video captures

a set of service behaviors $B = (b_1, b_2, \ldots, b_n)$, where n denotes the total number of distinct behaviors. Each $b_j \in B$ corresponds to a specific service behavior (e.g., five-finger guidance, hands offering, or maintaining good posture).

3.2 Problem Descriptions

To evaluate the performance of service behavior identification across n categories, this paper utilizes a confusion matrix $\mathcal{C}_{n\times n} = [c_{x,y}]_{n\times n}$. Each element $c_{x,y} \in \mathcal{C}$ represents the number of samples, where the true label is the category $b_x \in B$ and the predicted category is $b_y \in B$.

For each video segment $v_i \in V$, if the actual category of the segment v_i is b_x and the predicted category is b_y, the corresponding value of $c_{x,y}$ in the confusion matrix $\mathcal{C}_{n\times n}$ is updated as follows:

$$c_{x,y} = c_{x,y} + 1. \tag{1}$$

Let TP_j^M, FP_j^M and FN_j^M denote the true positive, false positive, and false negative, respectively, for the category b_j predicted using the mechanism M. These values can be calculated using the following equations:

$$TP_j^M = c_{j,j}, \tag{2}$$

$$FP_j^M = \sum_{x=1, x\neq j}^{n} c_{x,j}, \tag{3}$$

$$FN_j^M = \sum_{y=1, y\neq j}^{n} c_{j,y}, \tag{4}$$

where n denotes the total number of categories.

Let TP^M, FP^M and FN^M denote the true positive, false positive, and false negative for all service behaviors B using mechanism M, respectively. These values can be calculated using the following equations.

$$TP^M = \sum_{j=1}^{n} TP_j^M, \tag{5}$$

$$FP^M = \sum_{j=1}^{n} FP_j^M, \tag{6}$$

$$FN^M = \sum_{j=1}^{n} FN_j^M. \tag{7}$$

Let P^M, R^M and $F1^M$ denote the precision, recall, and F1-score for mechanism M, respectively. These values can be calculated using the following equations.

$$P^M = \frac{TP^M}{TP^M + FP^M}, \tag{8}$$

$$R^M = \frac{TP^M}{TP^M + FN^M}, \tag{9}$$

$$F1^M = \frac{2 * P^M * R^M}{P^M + R^M}. \tag{10}$$

3.3 Objective

Let Ω denote a set of potential mechanisms that are utilized to identify the specific service behavior. The primary goal of this paper is to find the optimal mechanism, denoted by $\mathbb{M}$, which satisfies Eq. (11).

Objective Function:

$$\mathbb{M} = \arg\max_{M\in\Omega} F1^M. \tag{11}$$

To achieve the above objective function, it is essential to satisfy the following constraints. Let $\alpha_{i,k}$ denote whether the segment v_i contains behavior b_k. That is

$$\alpha_{i,k} = \begin{cases} 1, & if\ v_i\ contains\ behavior\ b_k, \\ 0, & otherwise. \end{cases}$$

The following Category Constraint, which is described in Eq. (12), ensures that each audit segment $v_i \in V$ contains at most single service behavior $b_k \in B$.

1) Category Constraint

$$\forall v_i \subseteq V, such\ that \sum_{k=1}^{n} \alpha_{i,k} \leq 1. \tag{12}$$

The following Spatial-Temporal constraint described in Eq. (13), ensures that the same behavior b_k remains consistent across videos v_i and v_j within specific spatial and temporal bounds, accounting for variations in movement amplitude and action speed. Despite differences, both videos represent the same behavior.

2) Spatial-Temporal Constraint

$$S_i(t) = \gamma_{i,j} \cdot S_j(\delta_{i,j} \cdot t), \tag{13}$$

where $S_i(\cdot)$ and $S_j(\cdot)$ denote the spatial features at a time slot for v_i and v_j, respectively. The factor $\gamma_{i,j}$ denotes the difference in the amplitude of movements between v_i and v_j, while $\delta_{i,j}$ denotes a time-scaling factor.

The following Intra-Class similarity and Inter-Class Dissimilarity Constraint, described in Eqs. (14) and (15), ensures that different behaviors b_k and $b_l (k \neq l)$ are represented by significantly distinct feature vectors. Let $\phi(v_i \cdot \alpha_{i,k})$ and $\phi(v_j \cdot \alpha_{j,l})$ denote the high-dimensional feature vectors of v_i, belonging to the class b_k, and v_j, belong to the class b_l, respectively.

3) Dissimilarity Constraint

A. Intra-Class Similarity

$$\left\| \phi(v_i \cdot \alpha_{i,k}) - \phi(v_j \cdot \alpha_{j,k}) \right\| \leq \epsilon_k, \tag{14}$$

where $\|\cdot\|$ denotes an appropriate norm that measures the distance between feature vectors v_i and v_j, and ϵ_k controls the tolerance of intra-class similarity.

B. Inter-Class Dissimilarity

$$\left\|\phi(v_i \cdot \alpha_{i,k}) - \phi(v_j \cdot \alpha_{j,l})\right\| \geq \xi_{k,l}, \tag{15}$$

where $\xi_{k,l}$ specifies the tolerance for dissimilarity between different classes b_k and b_l.

4 The Proposed SMRT Mechanism

This paper introduces an innovative service behavior identification mechanism, called SMRT, designed to accurately identify service behavior for mystery shopping audit videos. SMRT focuses on actions like five-finger guidance, hand offerings, and proper posture. As shown in Figure 1, SMRT operates in two phases: coarse-grain and fine-grain identification. The coarse-grain phase detects 'person' objects in video frames, filtering out irrelevant ones, while the fine-grain phase extracts key behavioral features from the remaining frames. These frames are grouped into fixed-duration segments and analyzed using a Transformer model to capture temporal features. This integrated approach ensures precise identification and classification of service behaviors for comprehensive evaluations.

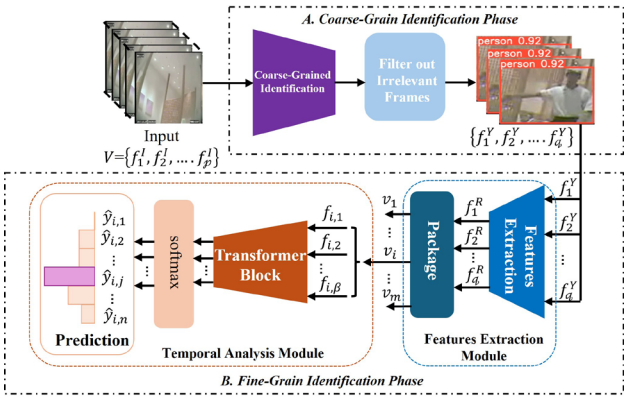


Figure 1. The process of the proposed SMRT mechanism

4.1 Coarse-Grained Identification Phase

This phase aims to identify 'person' objects in each frame of the mystery shopping audit video. As shown in Figure 2, frames without 'person' objects are filtered out, focusing the analysis on relevant frames and significantly reducing data volume. This enhances computational efficiency and accelerates processing, enabling SMRT to improve both detection speed and accuracy while optimizing resource utilization.

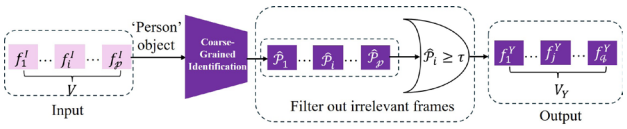


Figure 2. The process of coarse-grained identification

Formally, consider a mystery shopping audit video V, consisting of $\mathcal{p}$ continuous frames, represented as $V = (f_1^I, f_2^I, \ldots, f_{\mathcal{p}}^I)$, where $f_i^I \in V$ denotes i-th frame. The proposed SMRT processes each frame f_i^I for predicting the presence of a 'person' object. Let $\hat{\mathcal{P}}_i$ represent the confidence score provided by this phase for the 'person' object in the frame f_i^I, indicating the likelihood that a 'person' object is present. That is

$$\hat{\mathcal{P}}_i = Y(f_i^I), \tag{17}$$

where $Y(\cdot)$ denotes the coarse-grained identification model.

During the training phase, let $\mathcal{P}_i$ denote the true label for the frame f_i^I, where $\mathcal{P}_i = 1$ if a 'person' object is present in the frame f_i^I, and otherwise $\mathcal{P}_i = 0$. Let $\mathcal{L}(\mathcal{P}_i, \hat{\mathcal{P}}_i)$ denote the classification loss function for the 'person' object. The value of $\mathcal{L}(\mathcal{P}_i, \hat{\mathcal{P}}_i)$ can be calculated by Eq. (18).

$$\mathcal{L}(\mathcal{P}_i, \hat{\mathcal{P}}_i) = -(\mathcal{P}_i \log \hat{\mathcal{P}}_i + (1 - \mathcal{P}_i) \log (1 - \hat{\mathcal{P}}_i)). \tag{18}$$

During the inference phase, SMRT determines whether each frame f_i^I contains a 'person' object based on the confidence score $\hat{\mathcal{P}}_i$.

Assume that τ denotes a threshold, which decides if the frame f_i^I should be kept for further analysis. Let V_Y denote the filtered subset of the original video V. V_Y consists of frames likely containing a 'person' object. It is defined as Eq. (19).

$$V_Y = \{f_i^I \in V \mid \hat{\mathcal{P}}_i \geq \tau\}. \tag{19}$$

By isolating frames based on the confidence score $\hat{\mathcal{P}}_i$ and the threshold τ, this approach focuses the analysis on the most relevant parts of the video while significantly reducing the data volume for subsequent detailed analysis. This efficient filtering mechanism not only enhances the accuracy of the proposed SMRT in identifying frames with service provider behaviors but also improves computational efficiency, making the system more robust and scalable.

4.2 Fine-Grained Identification Phase

In this phase, the remaining frames in V_Y undergo detailed processing in two steps: Feature Extraction, and Temporal Analysis. The feature extraction aims to identify key behavioral attributes from each frame in V_Y. These frames are reassembled into fixed-duration segments. Temporal analysis employs a Transformer model to capture temporal features of these behaviors across the sequence of fixed-duration video segments. This enables precise identification and classification of specific service behaviors.

4.2.1 Features Extraction Module

This step aims to extract key behavioral features from each frame in V_Y. Assume that V_Y consists of q frames, represented as $V_Y = \{f_1^Y, f_2^Y, \ldots, f_q^Y\}$, where each frame $f_j^Y \in V_Y$ contains a 'person' object detected. This can be represented as Eq. (20).

$$f_j^R = R(f_j^Y), \tag{20}$$

where $R(\cdot)$ denotes the features extraction model, and f_j^R is the resulting feature vector for the j-th frame f_j^Y. After processing each frame in V_Y, a set of feature vectors is obtained, denoted by $V_R = \{f_1^R, f_2^R, \ldots, f_q^R\}$. This collection V_R represents the extracted key behavioral features for all q frames in V_Y.

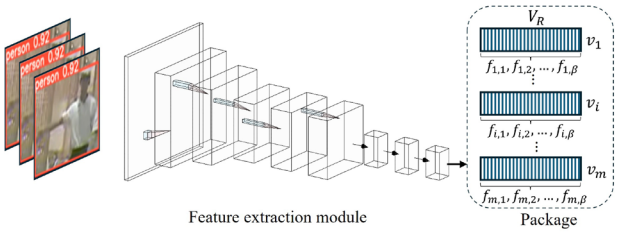


Figure 3. The process of the feature extraction module

As shown in Figure 3, to capture the temporal relationships between these frames, the feature vectors in V_R, are grouped into segments of a fixed duration, each containing β frames. This segmentation allows for effective temporal analysis using the Transformer model. The segmented set V_R is defined as $V_R = \{v_1, v_2, \ldots, v_m\}$, where each segment v_i comprises β feature vectors. These feature vectors within each segment v_i are denoted by $v_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,\beta}\}$ where each feature vector $f_{i,j}$ denotes that the j-th frame within i-th segment v_i, corresponding to the $((i-1)\times\beta + j)$-th feature vector from the V_R. That is

$$f_{i,j} = f_{(i-1)\times\beta+j}^R. \tag{21}$$

This structured segmentation process prepares the data for the Transformer's temporal analysis, allowing it to effectively capture and model the relationships and dynamics across the sequence of video frames.

4.2.2 Temporal Analysis Module

This step uses a Transformer to analyze the temporal features of fixed-duration segments, enabling precise identification of service behaviors. By leveraging its self-attention mechanism, the Transformer captures each segment of β frames to understand the temporal dynamics and dependencies within each segment v_i. This analysis allows the model to discern patterns and relationships both within and across segments, identifying key actions and interactions that constitute various service behaviors. Its ability to weigh the importance of each frame ensures accurate classification of both simple gestures and complex service interactions. The resulting predictions provide detailed insights into service quality, supporting performance assessment and training.

As shown in Figure 4, the process begins with each fixed-duration $v_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,\beta}\}$. Each v_i is first transformed through an input embedding layer, which converts the feature vectors into a suitable format for the Transformer model. This embedding layer yields $E_i = \{\varepsilon_{i,1}, \varepsilon_{i,2}, \ldots, \varepsilon_{i,\beta}\}$, where $\varepsilon_{i,j} \in E_i$ denotes the embedded representation of the j-th frame in the i-th segment.

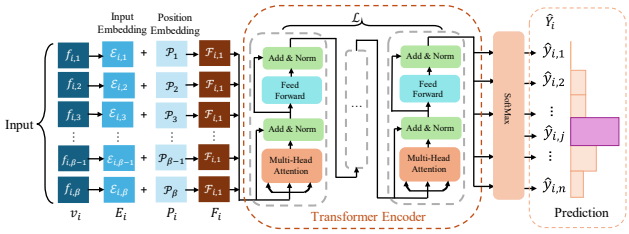


Figure 4. The process of the temporal analysis module

To incorporate positional information, which is important for the model to understand the order of the frames, position embedding, $P_i = \{\mathcal{P}_{i,1}, \mathcal{P}_{i,2}, \ldots, \mathcal{P}_{i,\beta}\}$ are added to the input embeddings. This combination results in $F_i = \{\mathcal{F}_{i,1}, \mathcal{F}_{i,2}, \ldots, \mathcal{F}_{i,\beta}\}$, where each $\mathcal{F}_{i,j}$ can be derived from Eq. (22).

$$\mathcal{F}_{i,j} = \varepsilon_{i,j} + \mathcal{P}_{i,j}. \tag{22}$$

Then, F_i is fed into the Transformer Encoder, which consists of $\mathcal{L}$ layers. Each layer of the Transformer Encoder comprises a multi-head attention mechanism followed by a feed-forward network. The multi-head attention mechanism is important for capturing different aspects of the relationships between frames within each segment. It operates on the input F_i using h heads to perform parallel attention operations. For the k-th attention head, let $\mathbb{Q}_{i,k}$, $\mathbb{K}_{i,k}$, and $\mathbb{V}_{i,k}$ denote the query, key, and value matrices of the input F_i, respectively. Let $W_{i,k}^{\mathbb{Q}}$, $W_{i,k}^{\mathbb{K}}$ and $W_{i,k}^{\mathbb{V}}$ denote the weights matrices of the $\mathbb{Q}_{i,k}$, $\mathbb{K}_{i,k}$, and $\mathbb{V}_{i,k}$, respectively. The values of $\mathbb{Q}_{i,k}$, $\mathbb{K}_{i,k}$, and $\mathbb{V}_{i,k}$ can be derived from Eqs. (23) to (25), respectively.

$$\mathbb{Q}_{i,k} = F_i \times W_{i,k}^{\mathbb{Q}}, \tag{23}$$

$$\mathbb{K}_{i,k} = F_i \times W_{i,k}^{\mathbb{K}}, \tag{24}$$

$$\mathbb{V}_{i,k} = F_i \times W_{i,k}^{\mathbb{V}}. \tag{25}$$

Each head then computes the attention scores using a scaled dot-product mechanism. Let $\mathbb{Q}_{i,k}$ denote the output of the k-th head, calculated by Eq. (26).

$$\mathbb{O}_{i,k} = SoftMax\left(\frac{\mathbb{Q}_{i,k}\mathbb{K}_{i,k}^{T}}{\sqrt{d_{\mathbb{K}}}}\right) * \mathbb{V}_{i,k}, \tag{26}$$

where $d_{\mathbb{K}}$ denotes the dimensionality of the key vectors, and the SoftMax function is applied to normalize the attention scores.

Instead of relying on a single-head attention mechanism, the Transformer employs multi-head attention to capture different aspects of the relationships between frames. The outputs from each attention head are concatenated and projected back into the original dimensionality. That is

$$\mathbb{A}_i = concat\left(\mathbb{O}_{i,1}, \mathbb{O}_{i,2}, \ldots, \mathbb{O}_{i,h}\right) \times W_i, \quad \textbf{(27)}$$

where $\mathbb{A}_i$ denotes the combined output of the multi-head attention for the first layer, W_i is a learned weight matrix, and h denotes the number of attention heads.

After the multi-head attention mechanism, each layer of the Transformer Encoder includes a feed-forward network (FFN) applied independently to each position. That is

$$\mathbb{O}_i = FFN\left(\mathbb{A}_i\right) + \mathbb{A}_i, \quad \textbf{(28)}$$

where $\mathbb{O}_i$ denotes the output of the FFN for the first layer. The output $\mathbb{O}_i$ is then fed into the next layer of the Transformer Encoder. This process is repeated for all $\mathcal{L}$ layers using Eqs. (21) to (26).

After processing through all $\mathcal{L}$ layers, the final output of the Transformer model represents the input segment v_i. This output is then passed through a SoftMax layer to generate a probability distribution over the possible service behaviors. During training, the Transformer model is optimized using a loss function. Let $y_{i,j}$ be the true label for segment v_i, where $y_{i,j} = 1$ if v_i corresponds to behavior b_j, and $y_{i,j} = 0$ otherwise. Let $\hat{y}_{i,j}$ denote the predicted probability of behavior b_j for the segment v_i. Let $\mathcal{L}(y_{i,j}, \hat{y}_{i,j})$ denote loss function, defined as Eq. (29).

$$\mathcal{L}(y_{i,j}, \hat{y}_{i,j}) = -\sum_{i=1}^{m}\sum_{j=1}^{n} y_{i,j} log\hat{y}_{i,j}, \quad \textbf{(29)}$$

where m and n denote the numbers of segments and behavior categories, respectively. By minimizing this loss, the Transformer model learns to improve predictions, enhancing its ability to identify and classify service behaviors.

5 Performance Evaluation

This section evaluates the proposed SMRT against the existing Lightweight Double-feature Triple-scale motion Network (LDT-NET) [23] and TinyVIRAT mechanisms [24]. LDT-NET, a skeleton-based method, struggled with poor image quality in covert videos, leading to inaccuracies in skeleton detection. TinyVIRAT, designed for low-resolution action recognition, incurs high computational overhead. In contrast, the proposed SMRT uses YOLO [9] for efficient 'person' detection, reducing computational load. It then employs ResNet-50 [10] to extract action features and a Transformer model for temporal analysis, improving accuracy in behavior recognition for services.

5.1 Dataset

The evaluation utilizes a custom dataset known as the Mystery Shopping Dataset, which includes 531 video clips capturing service-related activities—313 with standard service actions and 218 with non-standard actions. The dataset is split into 80% for training and 20% for testing. Notably, each frame in the videos is annotated, providing detailed frame-level information that enhances the model's accuracy in recognizing and distinguishing service actions.

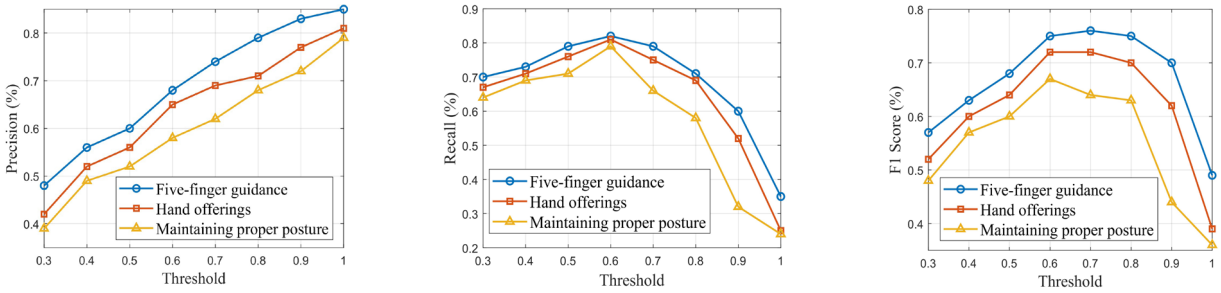
5.2 Simulation Results

Figure 5 illustrates the performance of the proposed SMRT across different behaviors (five-finger guidance, hands offering, and maintaining proper posture) in terms of precision, recall, and F1-Score under thresholds ranging from 0.3 to 1. Precision consistently increases with the threshold. The reason is that a higher threshold makes the model more confident in classifying instances as positive, thereby reducing false positives and boosting precision. In contrast, recall and F1-Score initially rise but then decrease as the threshold rises. Recall measures the model's ability to capture all relevant instances. At lower thresholds, the model tends to classify more instances as positive, including many true positives, which boosts recall. However, as the threshold rises, the model becomes more selective, reducing the number of true positives and causing recall to drop. F1-Score improves when both precision and recall are balanced but further increases in the threshold lead to a decline in F1-Score, as the loss in recall outweighs the gains in precision. Thus, careful threshold selection is crucial for optimizing SMRT's effectiveness in accurately recognizing and scoring behaviors.

Figure 6 compares SMRT with LDT-NET and TinyVIRAT in terms of precision, recall, and F1-Score as training videos increase from 100 to 400. Metrics improve with more training data due to enhanced learning and generalization. SMRT consistently outperforms both models, leveraging YOLO for object detection and ResNet-50 with Transformer models for spatial and temporal feature analysis. This combination enables SMRT to effectively capture and predict complex service behaviors.

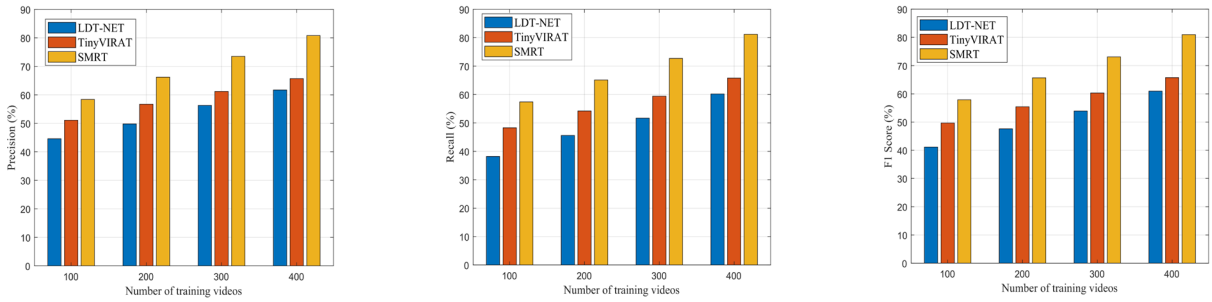
Furthermore, Figure 7 uses the Friedman test to compare the F1-Score distributions of LDT-NET, TinyVIRAT, and SMRT. The box plots show that SMRT exhibits a higher median F1-Score and more consistent distribution, indicating robustness and reliability. In contrast, LDT-NET and TinyVIRAT show a lower median F1-Score, suggesting less stability. The Friedman test results (chi-square = 8, p-value = 0.01832) confirm that these differences are statistically significant, supporting SMRT's superior performance over the other methods.

Table 1 presents an ablation study on the Mystery Shopping Dataset to evaluate the impact of YOLOv8, ResNet-50, and the Transformer encoder on SMRT's performance. The baseline model using only YOLOv8 for 'person' detection, achieves an F1-Score of 0.498. Adding ResNet-50 increases the F1-Score to 0.675, highlighting its role in improving feature extraction. Incorporating the Transformer encoder further boosts the F1-Score to 0.81, demonstrating its effectiveness in capturing temporal dynamics and contextual information.



(a) Precision (b) Recall (c) F1-Score

Figure 5. Comparison analysis of different behaviors in terms of precision, recall, and F1-Score



(a) Precision (b) Recall (c) F1-Score

Figure 6. Comparison analysis of different mechanisms in terms of precision, recall, and F1-Score

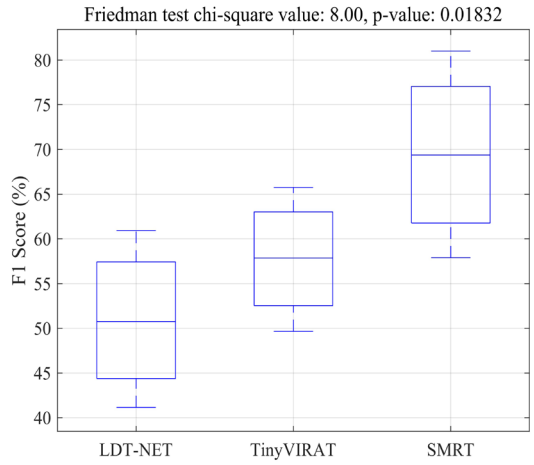


Figure 7. Comparison analysis of different mechanisms using the Friedman test

Table 1. The ablation study of the proposed SMRT

Method	F1-Score
YOLOv8	0.498
YOLO+ResNet-50	0.675
SMRT	0.81

6 Conclusion

This paper presents SMRT, an innovative service behavior identification mechanism for recognizing service provider actions in mystery shopping audit videos. SMRT effectively detects 'person' objects in the coarse-grain phase, even in blurred videos, and extracts key behavioral features in the fine-grain phase, using a Transformer model for temporal analysis. Experimental results demonstrate that SMRT excels in both accuracy and processing efficiency for behavior recognition tasks. Future work will focus on enhancing real-time processing to enable instant detection and analysis of service behaviors in video streams for real-time monitoring.

Acknowledgments

This work was supported in part by the Smart Home and Applied Industry Innovation Team, Higher Education Research Program Project under Grant No. 2022AH010067, Anhui Provincial Natural Science Foundation under Grant No. 2408085MF177, Anhui Outstanding Youth Fund Project under Grant No. 2022AH030109, and Open Foundation of Anhui Engineering Research Center of Intelligent Perception and Elderly Care under Grant No. 2022OPA02.

References

[1] D. Neimark, O. Bar, M. Zohar, D. Asselmann, Video transformer network, *IEEE/CVF International Conference on Computer Vision*, Montreal, BC, Canada, 2021, pp. 3156-3165. https://doi.org/10.1109/ICCVW54120.2021.00355

[2] P. Saini, K. Kumar, S. Kashid, A. Saini, A. Negi, Video summarization using deep learning techniques: a detailed analysis and investigation, *Artificial Intelligence Review*,

Vol. 56, No. 11, pp. 12347-12385, November, 2023. https://doi.org/10.1007/s10462-023-10444-0

[3] Ragedhaksha, Darshini, Shahil, J. Arunnehru, Deep learning-based real-world object detection and improved anomaly detection for surveillance videos, *Materials Today: Proceedings*, Vol. 80, pp. 2911-2916, April 2023. https://doi.org/10.1016/j.matpr.2021.07.064

[4] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, June, 2017. https://doi.ieeecomputersociety.org/10.1109/TPAMI.2016.2577031

[5] X. Deng, J. Liu, C. Peng, Y. Wang, Using improved YOLOv5 model to detect volume for logs in log farms, *Journal of Internet Technology*, Vol. 24, No. 7, pp. 1403-1413, December, 2023. https://doi.org/10.53106/160792642023122407002

[6] C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, 2023, pp. 7464-7475. https://doi.org/10.1109/CVPR52729.2023.00721

[7] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, D. Tao, A survey on vision transformer, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 1, pp. 87-110, January 2023. https://doi.org/10.1109/TPAMI.2022.3152247

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016, pp. 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[9] G. Jocher, A. Chaurasia, J. Qiu, *Yolo by ultralytics*, Code repository, 2023.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[11] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, *Annual Conference on Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014, pp. 568-576.

[12] S. Khan, A. Hassan, F. Hussain, A. Perwaiz, F. Riaz, M. Alsabaan, W. Abdul, Enhanced spatial stream of two-stream network using optical flow for human action recognition, *Applied Sciences*, Vol. 13, No. 14, Article No. 8003, July, 2023. https://doi.org/10.3390/app13148003

[13] Z. Wang, H. Lu, J. Jin, K. Hu, Human action recognition based on improved two-stream convolution network, *Applied Sciences*, Vol. 12, No. 12, Article No. 5784, June, 2022. https://doi.org/10.3390/app12125784

[14] A. Zhou, Y. Ma, W. Ji, M. Zong, P. Yang, M. Wu, M. Liu, Multi-head attention-based two-stream EfficientNet for action recognition, *Multimedia Systems*, Vol. 29, No. 2, pp. 487-498, April, 2023. https://doi.org/10.1007/s00530-022-00961-3

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, *IEEE International Conference on Computer Vision*, Santiago, Chile, 2015, pp. 4489-4497. https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.510

[16] Y. Zhou, X. Sun, Z. J. Zha, W. Zeng, Mict: Mixed 3d/2d convolutional tube for human action recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 449–458. https://doi.org/10.1109/CVPR.2018.00054

[17] H. Zhao, J. Liu, W. Wang, Research on human behavior recognition in video based on 3DCCA, *Multimedia Tools and Applications*, Vol. 82, No. 13, pp. 20251-20268, May, 2023. https://doi.org/10.1007/s11042-023-14355-8

[18] J. Y. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: Deep networks for video classification, *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, USA, 2015, pp. 4694-4702. https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7299101

[19] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, C. G. Snoek, VideoLSTM convolves, attends and flows for action recognition, *Computer Vision Image Understanding*, Vol. 166, pp. 41–50, Janaury, 2018. https://doi.org/10.1016/j.cviu.2017.10.011

[20] C. Dai, X. Liu, J. Lai, Human action recognition using two-stream attention based LSTM networks, *Applied Soft Computing*, Vol. 86, Article No. 105820, Janaury, 2020. https://doi.org/10.1016/j.asoc.2019.105820

[21] S. Yan, Y. Xiong, D. Lin, Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, *AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018, pp. 7444-7452.

[22] L. Wu, C. Zhang, Y. Zou, SpatioTemporal focus for skeleton-based action recognition, *Pattern Recognition*, Vol. 136, Article No. 109231, April, 2023. https://doi.org/10.1016/j.patcog.2022.109231

[23] M. Yin, S. He, T. A. Soomro, H. Yuan, Efficient skeleton-based action recognition via multi-stream depthwise separable convolutional neural network, *Expert Systems with Applications*, Vol. 226, Article No. 120080, Septemper, 2023. https://doi.org/10.1016/j.eswa.2023.120080

[24] U. Demir, Y. S. Rawat, M. Shanh, TinyVIRAT: Low-resolution Video Action Recognition, *International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 7387-7394. https://doi.ieeecomputersociety.org/10.1109/ICPR48806.2021.9412541

Biographies

Qiaoyun Zhang received her Ph.D degree in Computer Science and Information Engineering from Tamkang University. She is currently a Lecturer with the School of Artificial Intelligence, Chuzhou University, Chuzhou, Anhui, China. Her current research interests focus on artificial intelligence.

Shih-Yang Yang received his Ph.D degree in Computer Science and Information Engineering from Tamkang University, in January 2008. Since January 2021, he is an associate professor with the Department of Media Arts at Kang-Ning University. His research interests include parallel & distributed systems, web technology, and multimedia.

Yu Lin received her Ph.D. degree from the Graduate Institute of Life Sciences, National Defense Medical Center, in July 2018. She is currently an assistant professor in the Department of Nursing at Kang-Ning University. Her research focuses on medical technology.

Huan-Chao Keh is currently a full professor in the Department of Computer Science and Information Engineering at Tamkang University, Taiwan. He has been the President of Tamkang University since August 1, 2018. His current research interests include Data Mining, Internet of Things, Artificial Intelligence and Clinical Medical Information Systems.

Diptendu Sinha Roy received the Ph.D. Eng. degree from Birla Institute of Technology, Mesra, India in 2010. He is currently a professor with the Department of Computer Science Engineering, National Institute of Technology Meghalaya, India. His current research interests include 5G, software reliability, and machine learning, big data.