Human Position and Object Motion based Spatio-Temporal Analysis for the Recognition of Human Shopping Actions

Nethravathi P. S[1]
Faculty, College of Computer and Information Sciences Srinivas University
Mangalore, India

Karuna Pandith[2]
Faculty, Department of Information Science & Engineering NMAM Institute of Technology
Nitte, Karnataka, India

Manjula Sanjay Koti[3]
Faculty, Dept. of MCA
Dayananda Sagar Academy of Technology & Management
Karnataka, India

Rajermani Thinakaran[4*]
Faculty of Data Science and Information Technology
INTI International University
Nilai, Negeri Sembilan, Malaysia

Sumathi Pawar[5]
Faculty, Department of Information Science & Engineering
NMAM Institute of Technology
Nitte, Karnataka, India

***Abstract*—Retailers have long sought ways to better understand their consumers' behavior in order to deliver a smooth and enjoyable shopping experience that draws more customers every day and, as a result, optimizes income. By combining various visual clues such as activities, gestures, and facial expressions, humans may fully grasp the behavior of others. However, due to inherent problems as well as extrinsic forced issues such as a shortage of publicly available information and unique environmental variables, empowering computer vision systems to provide it remains an ongoing problem (wild). In this paper, the authors focus on identifying human activity recognition in computer vision, which is the first and by far the most important cue in behavior analysis. To accomplish this, the authors present an approach by integrating human position and object motion in order to detect and classify tasks in both temporal and spatial analysis. On the MERL shopping dataset, the authors get state-of-the-art results and demonstrate the capabilities of the proposed technique.**

Keywords*—Deep convolutional neural networks; computer vision; object detection; object localization; temporal analysis; human shopping actions component

I. Introduction

For years, the computer vision industry has been working on recognizing human actions. Many essential applications require the ability to recognize diverse behaviors from video data, such as fight identification from surveillance footage, human-robot interaction, video streaming analysis for online streaming services, and home security surveillance. Action recognition's main purpose is to recognize human actions in a video frame. For video-sharing services like YouTube and Twitch, action recognition is indeed a must-have feature. It can decipher a video's content and determine whether or not it should be made public. This tool can assist in the filtering of potentially harmful videos, such as bomb-making methods, choking activities, and the use of hard narcotics [1-3].

However, in areas like retail and shopping, the impact has been minimal. Using such technology in this context has a number of advantages, including efficient monitoring, consumer behavioral analysis, targeted marketing, and so on. Retailers will benefit from increased efficiency and revenue, as well as a more convenient shopping experience for customers if these strategies are used. Furthermore, the share of the worldwide trade market that these technological solutions occupy might be deduced from the rapidly expanding demand for them. The actual worth of such Artificial Intelligence (AI) based solutions relating to the retail industry is expected to be about US$10 billion by 2025, according to research conducted by Grand View Research [2].

Applying AI, and particularly machine learning approaches, to the shopping sector is still difficult, due to the insufficiency of data, primarily because of security issues, expensive labeling, as well as the need to stay proprietary where data is gathered. In spite of the datasets being publicly available to the researchers, (e.g., The MERL shopping dataset [4]), applying deep learning techniques to those is difficult as the external challenges posed by distinctive environmental factors like camera view angle, quality of the video, interrelations between the goods and the customers, and high obstruction. However, success of the existing deep learning algorithms can be attributed partially to the utilization of largely available public datasets like ImageNet [5], UCF101 [6], or [4], which allow sophisticated methods with numerous variables to be optimally trained. The completed actions are the major visual clue in understanding human behavior, which when paired with additional indicators like facial expressions tracked over time can also provide detailed behavioral knowledge [1, 7, 8,]. To describe human actions, the initial efforts on action recognition adopted Three-dimensional (3D) models [9, 10]. However, creating a 3D model using videos is time-consuming and costly.

As a consequence, people rather employ global or local representations for action recognition. These methods are known as representation-based methods. Currently, deep networks-based algorithms have indeed been able to attain promising results in action detection, because of the fast evolution of graphics processing units (GPU). The task of activity detection and recognition from shopping surveillance footage inputs is used as the primary topic of this paper. The current study only focuses on categorizing the clipped video into the given activities during recognition, whereas categorization is practiced to a continuous video sequence of multiple activities during detection; i.e., the temporal location at the beginning and length of the actions are also unknown and desired [11].

The main contributions of this paper are

- For a less explored camera view angle, we present an innovative strategy powered by Generative Adversarial Networks (GANs) which uses partial body position in the lack of exact joint locations (top view).
- Using the proposed novel method along with the standard transfer learning, we train and test on the MERL shopping dataset, which has different challenges such as camera angle view, classifications of activity, and limited data for training.
- The authors propose a simple but successful technique for using our action-identifying network as an action detector that identifies and classifies action locations in real-time. This method divides the difficult detection task into identification and detection modules and uses a two-stream network to combine diverse sources in semantic space.
- The authors successfully combine two independent sets of features, one for recognizing or detecting the action, namely human body posture (incomplete) and object-of-interest motion, to direct greater network resources and attention to the most significant signals while ignoring the less important ones. This is performed through the use of self-attention, in which the video's relevant spatial regions are connected to other regions in adjacent frames for enhanced accuracy and/or precision.

The remaining sections of the paper are organized as follows: Section II - consists of the extensive literature survey; Section III - is the detailed explanation of the methodology used in the study; Section IV - is a detailed presentation of the results arrived at along with comparisons. We conclude the paper by providing future directions.

II. Literature Survey

Extraction through feature engineering algorithms that can effectively recognize and depict motion in the input pattern of image frames is the focus of this research. Many approaches of mixing optical flow (OF) and feature matching [12] have been introduced since the early phases of the space-time pyramid [13] until recent years, with the most current ones attempting to replace feature extractors using deep neural networks. However, predicting the flow of optical from succeeding video frames has proven to be quite efficient. The literature on this subject contains a multitude of ways. Recently, there have been attempts to integrate these techniques with deep neural networks in order to achieve the best-of-both-worlds results, Horn et al. [14]. Many indigenous features have been facing this situation in recent years.

Pose estimation from single red, green, and blue (RGB) images has made substantial progress recently [6, 11, 15, 16, 17,], prompting its use as a high-level feature in movies to effectively represent diverse sorts of activities [18, 19]. Single image estimate is relatively reliable, even the tiniest mistake or disturbance in sequential posture estimation in videos is detrimental to activity interpretation utilizing existing futuristic techniques. With regards to missing joint locations in frames, as we will show factually, this occurs rather frequently irrespective of regular or low demanding settings. As a result, several techniques to utilize this vital information as an additional medium of data in combination with other resources such as optical flow and raw input frame have been presented. There are many alternative ways to take advantage of body posture characteristics while disregarding the faults. Indigenous approaches to cipher consecutive pose data into photos for classifying the actions are defined by some. There are numerous approaches to efficiently incorporate posture estimation into activity understanding in case of good pose prediction (for instance, 3D pose employing motion capture skeletal sensors or depth camera) [20]. However, because of a crowd, low range depth and occlusion, and costs, the use of depth cameras or signal fusion techniques is not viable in the shopping environment.

In computer vision, human-object interaction has long been a concern. However, the majority of the considered interactions revolve around sports [21], cooking [22], or ordinary activities [6, 9], which can sometimes be categorized from single photos [21]. Kim et al. [23] present an effective method for recognizing object-based activities. For action recognition from security cameras, the researchers make use of the graph neural networks for merging the object and human pose data. Their method, however, is largely dependent on the quality of the input posture data.

Multi-stream Convolutional Neural Networks (CNNs) have recently been popular for combining diverse modalities of data before generating decisions [4, 16]. It can be seen in most of the presented systems that one stream is dedicated to temporal understanding (usually utilizing a labeled training data and computed algorithm related to flow of optics [17]), whereas another is exclusively for diving into spatial features in the image.

Many academics have been attributed to the success of Two-dimensional (2D) CNNs in image interpretation to investigate the feasibility of doing so in videos. As a result, numerous ways to widen the convolution in time have been developed [7, 23, 24]. One of the issues with these techniques is that most of these are computationally costly, whilst moderately outperforming 2D CNN counterparts.

Several researchers have rethought their application, inventing sophisticated combinations of 2D and 3D CNNs to achieve the best possible outcomes [25, 26]. Nevertheless, the

work of exploiting pose data is still under-explored to our knowledge. Many other researchers integrate the two works and sort them concurrently [27]. Some techniques solely tackle the temporal detection element of the work, whilst others integrate both activities and sort them at the same time. Some are inspired by object identification techniques and use temporal region proposals [10, 18], while others are known as segmentation.

In the natural language processing (NLP) community, many unique strategies to replicate human attention in CNNs were first proposed on machine translation jobs. Some of them are used in sequential tasks to highlight key input frames [28], while others are used to simply focus on spatial regions of relevance [10, 9, 2]. Wang et al. [29] and combine them into one differentiable peripheral module. Moreover, there are works defining a representative capable of discovering the important areas or frames using sophisticated analytics like Reinforcement Learning (RL) [28]. The 3D modeling method was utilized extensively in early action recognition studies. The walker hierarchical model [9] uses a number of hierarchical levels to depict a person. To recognize pedestrians in a video, [10] employs linked cylinders and their progression. The kernel learned by CNN is visualized and found that the bottom layers learn low-level features while the upper layers learn high-level representations. This demonstrates that convolutional architecture may be utilized to extract features [29, 30].

Videos, unlike photographs, have a dynamic nature. Directly adding temporal features to a convolutional architecture is a typical approach of using deep networks for action recognition. To achieve this, Ji et al. [31] postulated the 3D CNN, which also uses 3D kernels to obtain both spatial and temporal information.

Many researchers have made contributions to the field of temporal information integration in CNNs. In the temporal domain, Ng et al. [30] discover that maximal pooling outperforms average and other pooling approaches. Karpathy et al. [32] present the slow fusion model, a new convolutional architecture that receives video clips and processes them via an identical set of layers (with common parameters) to provide outcomes for fully connected layers. The video description will then be generated from these completely connected layers. Tran et al. [33] combine the concepts of the visual geometry group (VGG) [9], Decaf [14], as well as the 3D CNN [29] to develop a 3-D graphics technique that can build 3D graphics out of 2D images (C3D), a generic video descriptor. They use the Sports-1M [32] dataset to train their network and extract video attributes from such a fully - connected layer. Learning temporal information from uncontrolled input is the purpose of a deep generative network [21, 26]. Xing Yan et al. [34] present a deep Dyn encoder to capture video dynamics, based on the linear dynamic system modeling approach proposed by Doretto et al. [16]. The Long Short-Term Memory (LSTM) autoencoder model is created using the LSTM [26] cell.

The encoder LSTM and the decoder LSTM make up this model. Goodfellow et al. [35] propose an adversarial network to address the training challenges in deep generative networks. The competition between a generative and a discriminative model is referred to as adversarial. Mathieu et al. [36] use the adversarial principle where a multi-scale convolutional network is trained and highlight the benefits of pooling in a generative model. In this body of work, many databases have been added to aid in the development and testing of algorithms [28, 6, 24]. Kinetics [8], which is called the ImageNet [13] of videos, is among the largest. Only a few of these have untrimmed films that can be detected [22, 27]. More crucially, as previously said, the exclusive characteristic of statistics related to retail is the insufficiency of data on which to test frameworks to address the distinctive difficulties. As per our knowledge, the MERL [4] dataset is the only one available for human shopping actions. All of these aspects are addressed using Human Position and Object Motion based spatiotemporal analysis and are tested on the Recognition of Human Shopping Actions. Further details regarding the same are explained in the subsequent sections.

III. METHODOLOGY

A GAN consists of two approaches: a generic model that is trained to comprehend the probability distribution of the input data and a discriminative model that seeks to distinguish real input samples from false ones [37]. Backpropagation is used to simultaneously train both models. GANs' success has been seen in their wide range of applications, which range from creative picture generation and super-resolution to semi-supervised classification. Using input pictures and noisy and imprecise joint heat maps, the authors propose a conditional GAN-based technique for regressing the precise location of six body joints. In the retail environment, we commonly encounter top-view (or near-top-view) recorded surveillance cameras, which exacerbate the difficulty of deciphering human activity by imposing extra occlusions of various body parts and components. Because of this distinct and demanding perspective, many deep learning professionals have avoided training on these sorts of input images. The posture estimator system that we implemented in our challenge suffers from the same problem, despite producing state-of-the-art results of different distinct activities that support traditional camera view aangles Fig. 1 depicts the challenges encountered by the given strategy in the present working dataset.

Fig. 1. Illustration of the Issues Faced due to the Camera Angles.

To improve posture estimate accuracy, we present the GAN structure. It also ensures that the positions of the joints of interest in the dataset are predicted more accurately. The generator is in charge of learning the conditional probability of the joint locations given to the current noisy heat map that is extracted and the input frame, which is a CNN with an architecture similar to Inception-v4 followed by a multi-layer perceptron (stacked). The same has been summarized in Fig. 2.

In our scenario, only six joints are considered essential in the current shopping environment scenario, which are both the left and right shoulders, elbows, and wrists. To begin, consider human posture as a probabilistic heat map indicating the areas of joints that require better attention. Next, treating it as a high-level feature of the object which is moving in the scene, rather than treating pose data as a separate source of information for defining an action, instead of general features extracted with a massive proportion of unnecessary background data to acquire motion data without considering any other specific motion representation, forcing the deep learning architecture for feature extraction with a greater focus on these spots in each frame of the input images. Then, by implementing the GAN fine-tuning step, we reduce even more noise from the heat maps obtained from the input related to the six key joints and obtain its precise location, which is given to our pose stream network as a replacement for the pose heat map channel. The working of the same has been described in Fig. 3.

Refer to (1)

$$y_i = \frac{1}{c(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \quad (1)$$

where, i and j are indices of the locations in frames over the entire sequence (space-time), f refers to embedded Gaussian mapping of the input pattern, and g is a linear mapping function. Here x_i refers to weight for a single location in space-time input, and j is the set of all other potential places. The Embedded Gaussian Function is the function we're using here. The number of modules inserted into the network for optimal performance has been empirically determined to be two. The LSTM's hidden state vectors are subjected to the second attention mechanism. It actually assigns a scalar weight to every input frame based on the network's learned relevance. These weights were adopted to the LSTM's hidden feature vector as in practice. At every time the first step (relating to the feature vector of the frame is multiplied by the weight value supplied by the temporal attention), the LSTM has a hidden state. This is done in practice where a single fully-connected layer on the LSTM's hidden states is trained. The first of the two modules, in particular, has played a key role in advising us on how to strengthen our approach. The placement of joints is connected with a higher weight, notably the six joints that make up a part of the posture model, as shown in the representations provided in Fig. 3. By substituting the exact heat map of joint locations with the ones from our generator network, there was even more attention forced which improved our outcomes. In this case, by reducing the noisy heat, we were able to stimulate the concentration of network re- sources with more assurance maps.

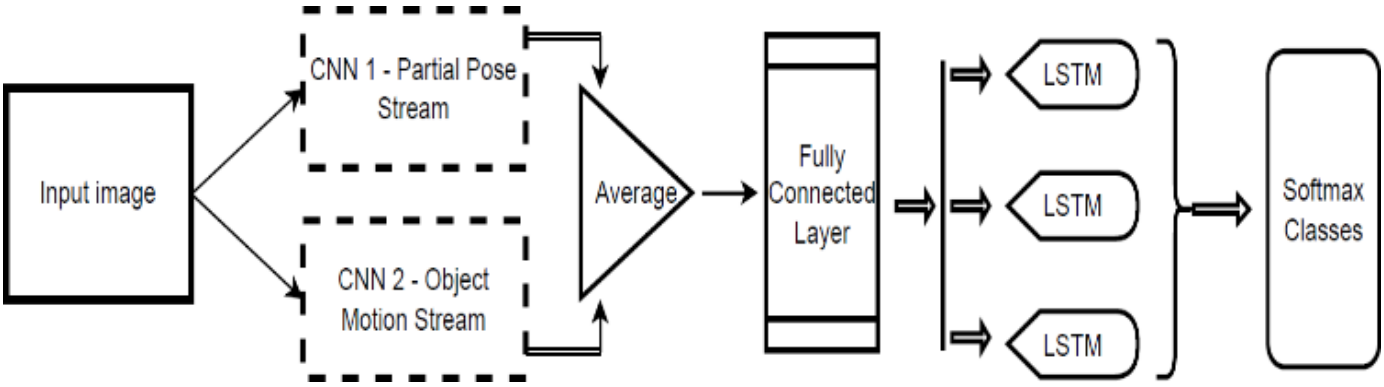

Fig. 2. An Overview of the Proposed GAN Structure.

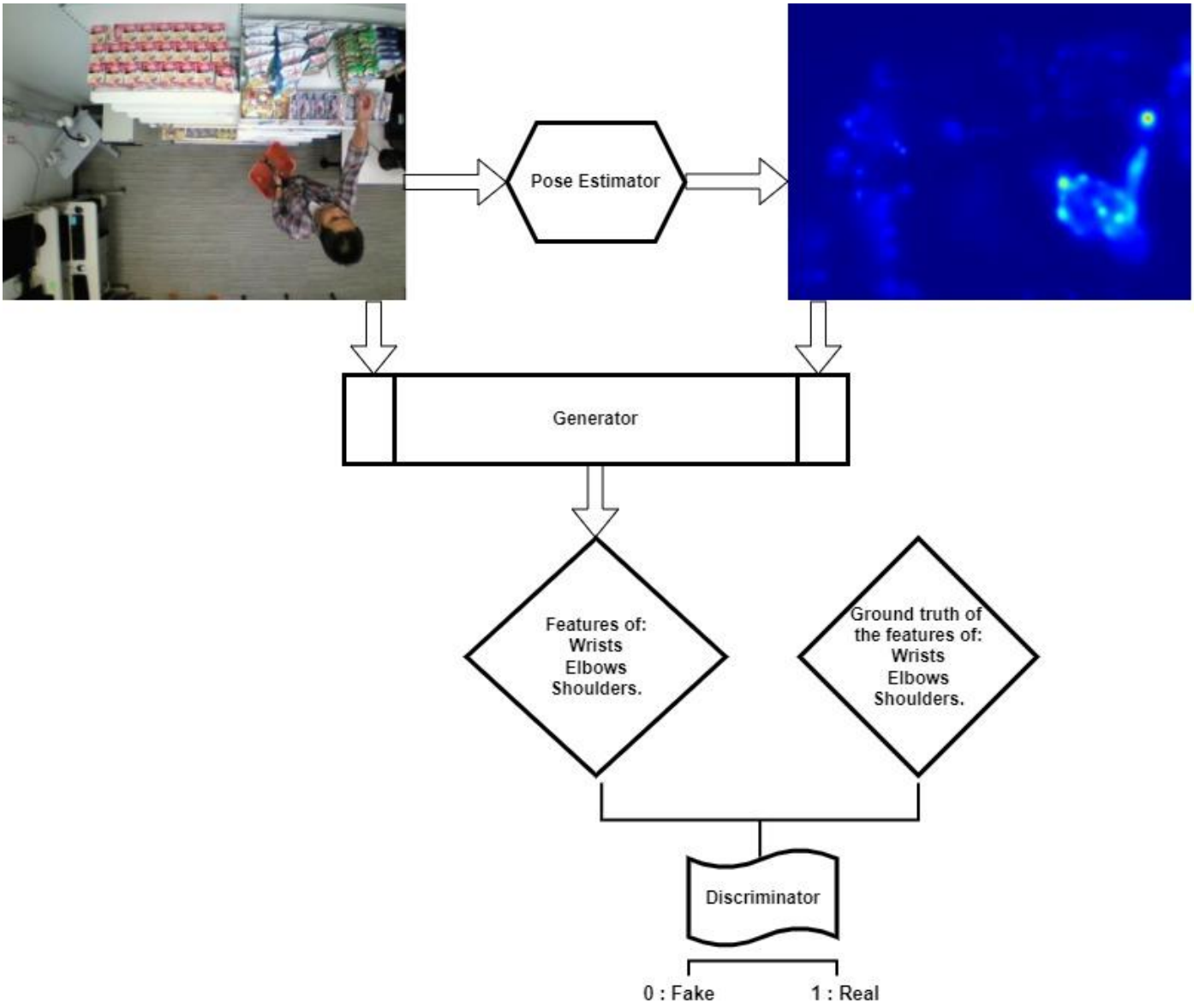

Fig. 3. A Summary of our GAN-Based Pose Fine-Tuning Method.

IV. Results and Analysis

A. Experimental Setup

The authors used PyTorch for the implementation as it is open source. The fine-tuned network is then used with the remainder of the network, with no further training, to retrieve earlier data regarding item placements. To construct the weights for the generator in GAN architecture, we manually annotated additional thousand frames which are sampled uniformly with six joint-f-interest (e.g., wrists, elbows, and shoulders) positions. After that, the generator is supervised, and they are pre-trained using the Mean Square Error loss function across the ground-truth joint locations. On the other hand, the discriminator was built up at random using the Xavier approach.

After the GAN training has converged, the generator is used individually along the remainder of the network without any more fine-tuning, as previously indicated. The pose and object mappings are created using binary maps in the following manner. The map has values of one in a circle of constant radius (here 10 pixels) around each joint center and 0 somewhere else for joint locations, and one in a rectangle area of fixed size (here 40 pixels) around the center of each identified item and zero otherwise for object locations. To reduce the Cross-Entropy loss over the Softmax class probability outputs, the entire recognition network is trained using gradient descent.

Adam is the optimizer that has been proven to perform better in terms of fast convergence in large-scale models like ours. This is achieved by making dynamic (or adaptive, as the authors call it) changes to the learning rate for each weight based on the gradient's higher and lower coefficients so far. The default hyper-parameter settings specified by the authors are = 0.001, 1 = 0.9, 2 = 0.999, and 1008.

It's difficult to train the spatiotemporal attention components in the intermediary layers of the two streams. These modules were given a contribution weight gamma (constant across the network) that was initially set to 0. The authors restart the training for a few epochs (9 to 11) once the main network has converged, starting with 0.1 and gradually increasing it to 1. Finally, because each stream has a similar design to the Inception v4 network, the authors initialize both streams using transfer learning from networks trained on COCO classification tasks. This allows us to drastically minimize training time; the full network training takes only 17-21 epochs to complete.

The detection sliding window and stride parameters must be fine-tuned in the final stage. A Brute Force search in two sets of values is used to accomplish this. As mentioned in the assessment section, the overall performance of each pair of window-stride size values is examined, and the top overall values are picked. Finally, because each stream has a similar architecture to the Inception v4 network, we initialize both streams using transfer learning from networks trained on COCO classification tasks.

This allows us to cut the training time in half; the total network training takes around 17-21 epochs. The detection sliding window and stride parameters must be fine-tuned in the final stage. A Brute Force search in two sets of values is used to accomplish this. As noted in the assessment section, the overall performance of each pair of window-stride size values is evaluated, and the top overall values are picked. Sliding windows range in length from 3 to 45 frames, with a stride of one to the length of the window. The PyTorch deep learning library is used to implement the entire training and inference process in Python.

B. MERL Dataset Results

The MERL dataset was produced in response to a lack of relevant datasets in retail settings. The six activities are "reach the shelf," "retract from the shelf," "hand in the shelf," "inspect the object," "inspect the shelf," and the backdrop (or no action) class. It was photographed using a roof camera to look like a real retail mall, but not in any detail. There are 42 people in this dataset who work as shoppers. One of the dataset's challenges is that participants must complete a sequence of tasks, which can be easily exploited as a well-behaved transition probability matrix to produce good results on this dataset at the cost of simplifying it as a well-behaved transition probability matrix. Table I is an example of this challenge. This supplies the network with useful prior knowledge that may be used during inference to improve recognition accuracy.

TABLE I. MERL DATASET, ACTIONS, AND ITS CORRESPONDING DISTRIBUTIONS

Actions	Reach	Retract	Hand in	Inp. Product	Insp. Shelf
Reach	0.0	63.8	34.1	0.7	1.4
Retract	21.2	0.0	0.8	49.53	28.47
Hand in	0.12	85.7	0.0	6.32	7.86
Inp. Product	61.08	3.1	0.84	0.0	34.98
Insp. Shelf	98.9	0.094	0.67	0.34	0.0

Nonetheless, one of the key benefits of our technique is that we may get cutting-edge findings without relying on this extensive historical knowledge. This makes sense since, in real-world solutions, powerful priors are difficult to come by and impractical, therefore they can't be used. As a result, the authors eliminated these biases by ensuring that the participants do not follow a straightforward sequence of activities during the testing.

Unlike prior approaches that reported on the MERL dataset, we present results for both recognition and detection. The former believes that the input films only include single action, but the latter receives an untrimmed video with numerous actions as input. Table II displays our MERL recognition results, and Table III compares our detection results to those previously published because the recognition results for this dataset haven't been disclosed before by any other method, this presents a new challenge for future research, allowing other algorithms to compare their detection accuracy. As we've seen and as our results show, the accuracy of detection should be near to the precision of recognition for a decently good detection approach. As compared to other approaches, we significantly have higher Intersection over Union (IoU) (average 0.77 compared to 0.5 as the maximum reported) over previous methods while attaining better detection results, too.

TABLE II. RESULTS OF PROPOSED METHOD ON MERL DATASET

Details	In Percentage
Overall recognition Accuracy	71.14
F1 @ 50	67.11
Frame wise accuracy	71.41

TABLE III. COMPARISON OF RESULTS WITH PROPOSED METHOD (MERL DATASET)

Methodology	F1 {IoU = 0.5}	Accuracy in Percentage (Frame-Wise)
Multi-stream bi-directional RNN [4]	65.4	76.3
Temporal Convolutional Networks	72.9	79.0
Two stream CNN	74.8	77.1
Proposed methodology	77.46	75.13

V. CONCLUSIONS

The authors have developed a framework for fine-grained activity recognition and detection in retail environments. Due to the short time between each action, considerable intra-class variance, and low inter-class variation, fine-grained detection is difficult; these difficulties contribute to the task's intrinsic complexity. Furthermore, the work is made more difficult by the extrinsic difficulty of a rare and unusual camera viewing angle. We developed a semi-supervised technique employing GAN to fine-tune posture estimate results when there is a discrepancy in the images present at different angles during training and prediction. We gave detailed experimental data to demonstrate the method's applicability in real-world scenarios, particularly in shopping contexts, which have their own set of characteristics and obstacles.

When direct estimation fails, we speculate that combining pose estimation with image sequences is one way to use the associated range of motion as background information for forecasting joint location in the next time step (e.g., occlusion). Furthermore, we believe that describing the attention process as a salient region localization job for an expert system might lower the computation cost of its training and prediction, given the recent success of deep reinforcement learning methodologies. Further attempts at face expression recognition and/or eye gaze prediction provide more data for deep customer behavior analysis.

REFERENCES

[1] K. Soomro, H. Idrees, and M. Shah. Online Localization and Prediction of Actions and Interactions. volume 41, Feb.2019.

[2] H. B. Zhang, Y. X. Zhang, B. Zhong, Q. Lei, L. Yang, J. X. Du, and D. S. Chen, "A comprehensive survey of vision-based human action recognition methods," Sensors, 19(5), pp. 1005, 2019.

[3] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in human action recognition: A survey," 2015, arXiv preprint arXiv:1501.05964.

[4] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1961–1970, June 2016.

[5] J. Deng, W. Dong, R. Socher, L. Li, and and. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, June 2009.

[6] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild.Dec. 2012. arXiv: 1212.0402.

[7] G. Tripathi, K. Singh, and D. K. Vishwakarma. Convolutional neural networks for crowd behaviour analysis: a survey. pages 1–24. Springer, 2018.

[8] S. Nigam, R. Singh, and A. K. Misra. A Review of Computational Approaches for Human Behavior Detection. May 2018.

[9] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial Transformer Networks. June. 2015. arXiv: 1506.02025.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. June 2017. arXiv: 1706.03762.

[11] R. Bai, Q. Zhaoy, S. Zhou, Y. Liz, X. Zhaox, and J. Wang. Continuous action recognition and segmentation in untrimmed videos. August 2018.

[12] H. Wang and C. Schmid. Action recognition with improved trajectories. In 2013 IEEE International Conference on Computer Vision, pages 3551–3558, Dec 2013.

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008.

[14] B. K. Horn and B. G. Schunck. Determining optical flow. volume 17, pages 185–203, Aug. 1981.

[15] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. Jan. 2016. arXiv: 1602.00134.

[16] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, Computer Vision ECCV 2016, pages 483–499, Cham, 2016. Springer International Publishing.

[17] R. A. Gler, N. Neverova, and I. Kokkinos. DensePose: Dense Human Pose Estimation In The Wild. Feb. 2018. arXiv:1802.00434.

[18] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d Pose Estimation and Action Recognition using Multitask Deep Learning. Feb. 2018. arXiv: 1802.09232.

[19] R. Poppe, "A survey on vision-based human action recognition," Image and vision computing, 28(6), pp. 976-990, 2010

[20] R. Vemulapalli, F. Arrate, and R. Chellappa. Human Action Recognition by Representing 3d Skeletons as Points in a Lie Group. pages 588–595, 2014.

[21] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 17–24, June 2010.

[22] M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, and B. Schiele. Recognizing Fine-Grained and Composite Activities using Hand-Centric Features and Script Data. volume 119, pages 346–373, Sept.2016. arXiv: 1502.06648.

[23] S. Kim, K. Yun, J. Park, and J. Y. Choi. Skeleton-based action recognition of people handling objects. 2019.

[24] R. Hou, C. Chen, and M. Shah. An end-to-end 3d convolutional neural network for action detection and segmentation in videos. 2017.

[25] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Tradeoffs in Video Classification. Dec. 2017. arXiv: 1712.04851.

[26] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. Dec. 2016. arXiv: 1612.01925.

[27] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal Action Detection With Structured Segment Networks.pages 2914–2923, 2017.

[28] Y. Rao, J. Lu, and J. Zhou. Attention-Aware Deep Reinforcement Learning for Video Face Recognition. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 3951–3960, Venice, Oct. 2017. IEEE.

[29] X.Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. IEEE, Jun 2018.

[30] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. "Beyond short snippets: Deep networks for video classification". In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 4694–4702.

[31] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. "3D convolutional neural networks for human action recognition". In: IEEE transactions on pattern analysis and machine intelligence 35.1 (2013), pp. 221–231.

[32] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Suk-thankar, and Li Fei-Fei. "Large-scale video classification with convolutional neural networks". In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014, pp. 1725–1732.

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks". In: Proceedings of the IEEE International Conference on Computer Vision. 2015, pp. 4489–4497.

[34] Xing Yan, Hong Chang, Shiguang Shan, and Xilin Chen. "Modeling video dynamics with deep dynencoder". In: European Conference on Computer Vision. Springer. 2014, pp. 215–230.

[35] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: Advances in neural information processing systems. 2014, pp. 2672–2680.

[36] Michael Mathieu, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error". In: arXiv preprint arXiv:1511.05440 (2015).

[37] Mogren, O. (2016). C-RNN-GAN: Continuous recurrent neural networks with adversarial training. arXiv preprint arXiv:1611.09904.