Research on Deep Learning-based Group Recognition

Songchen Xu *

Faculty of information technology, Beijing University of Technology, Beijing, 100124, China

* Corresponding Author Email: ericxusc09@emails.bjut.edu.cn

Abstract. The task of group recognition is a process of automatically recognizing and analyzing individuals or groups from a large population using computer vision and machine learning techniques. Currently, the main methods for group recognition using deep networks can be divided into four categories: methods based on convolutional neural networks (CNN), methods based on recurrent neural networks (RNN), combined CNN-RNN-based methods, as well as pose analysis methods. This article specifically analyzes these four categories of methods: CNN-based, RNN-based, and pose analysis, and summarizes the characteristics and functions of iterative deep learning models, graph convolutional network models, long short-term memory models, pose models, and skeleton models. In addition, a detailed introduction to commonly used datasets in group recognition is provided, and the main evaluation metrics are analyzed. Finally, important research directions in group recognition are discussed, followed by a conclusion of the entire article.

Keywords: group recognition, deep learning, convolutional neural network, recurrent neural network.

1. Introduction

With the advancement of deep learning and the wide application of group recognition in areas such as crowd management, security monitoring, and traffic flow analysis, group recognition based on deep learning has gradually become a hot research topic in today's society. It can help in real-time monitoring and analysis of crowd dynamics, provide decision support and security assurance, and also serve as an important data source for group behavior research and urban planning. Generally, traditional group recognition methods are broadly classified into two types, namely probabilistic graphical models and grammar models. The specific process includes data collection, pedestrian detection and tracking, feature extraction, feature representation and encoding, classifier training and testing, and crowd analysis. Among them, the representation and encoding of features are the most critical components that impede traditional methods, often being affected by lighting conditions and occlusion.

With the emergence of deep learning, these issues are being addressed, and group recognition based on deep learning exhibits better performance and stronger robustness. Deep learning-based technologies developed at an extremely rapid speed in recent years may be organized into the following four groups:

1) Group recognition based on convolutional neural networks (CNNs). Convolutional Neural Network (CNN) is a deep learning model mainly used for image and video processing tasks. It consists of multiple convolutional layers, pooling layers, and fully connected layers. With the characteristics of parameter sharing and sparse connections, CNN has been widely applied in the field of computer vision and has achieved excellent performance. Through end-to-end learning, it can automatically learn more discriminative feature representations from the data, thereby improving the performance and robustness of group recognition. Additionally, it can be combined with other techniques such as object detection and pose estimation to further enhance the effectiveness of group recognition. As technology continues to develop and application demands expand, CNN-based group recognition methods will continue to drive research and applications in related fields.

2) Group recognition based on recurrent neural networks (RNNs). Recurrent Neural Network (RNN) is a type of neural network model with recurrent connections. Compared to traditional feedforward neural networks, RNNs is capable of processing sequential data like text, audio, and time-series data that have temporal dependencies. Its core lies in the ability to remember previous network states through recurrent connections, enabling a stronger understanding of context. This

method can also handle time dependencies in sequential data, making it suitable for modeling and analysis of pedestrian trajectories, video sequences, etc. Moreover, it can be combined with other techniques, such as using CNNs to extract features from image frames, to further improve the effectiveness of group recognition. With continuous technological advancements and increasing application demands, RNN-based group recognition methods are expected to receive widespread research and application.

3) Combination of CNNs and RNNs. In this approach, the main purpose of combining CNN and RNN is to leverage the strengths of both models and create a more powerful model for handling data with spatial and temporal dependencies simultaneously. This combination is commonly used to process spatial and temporal features in sequential data, such as video data and speech data, to accomplish more complex group recognition tasks.

4) Other methods. In addition to the above-mentioned approaches, there are also other methods, such as those based on pose analysis. Pose analysis has multiple applications in group recognition. Group recognition refers to the analysis and recognition of collective behaviors, interactions, and attributes within a group of people. Pose analysis can provide information about individual poses and group dynamics, thus offering important features for group recognition. Its applications are extensive, including tasks such as human pose estimation, behavior analysis, group structure analysis, group activity recognition, and group emotion analysis. It can provide rich features and contextual information for a comprehensive understanding of group behaviors and interactions.

This paper analyzes and reviews the research progress and current status of deep learning-based group recognition methods discussed above. It provides a detailed introduction to the target datasets and detection results in the field of group recognition. Reasonable predictions are made regarding the possible future development directions in the field of group recognition.

2. Traditional model-based group recognition

Traditional model-based group recognition refers to the use of traditional algorithms, which is based on machine learning and the techniques in computer vision to process group behaviors and features. It can be divided into two types: probabilistic graphical models and grammar models.

2.1. Probabilistic graphical models

Probabilistic graphical models use graphs to represent probability distributions. These models utilize graph theory and combine it with probability theory to represent the dependencies between various objects and obtain the joint probability distribution of all variables in the model.

Probabilistic graphical models are formed by two main ingredients, which are nodes and edges. Nodes represent random elements, while edges represent the huge number of relationships between different elements. Based on the type and connectivity of edges, probabilistic graphical models might be classified into different groups: straight graph models and indirect graph models.

Probabilistic graphical models provide an intuitive and efficient way to represent relationships between variables and allow for probabilistic inference and uncertainty reasoning. Through probabilistic graphical models, unknown variables can be analyzed by using known information through prediction, classification, clustering, and other analyses. They are also helpful for designing and developing new models and simplifying mathematical expressions for complex inference and learning operations.

In 2009, Wei et al. combined the probabilistic summation framework with spatiotemporal interest points and applied it to the decision-making process of object recognition [1]. In the experiment, the extracted spatiotemporal interest points contained the main objects in the scene, such as athletes and balls in a sports scenario. Spatiotemporal features accurately describe the movements of people in the scene without the need for tracking and action segmentation, thereby providing precise object categories.

In 2013, Yin et al. proposed a Gaussian process-based method within a probabilistic framework to recognize the temporal activities of group members [2]. They also applied a robust multi-target tracking algorithm to track each individual in the entire image region. Then, based on the output positions of the trackers, they performed clustering on the new group. Through a trained Gaussian model, they achieved recognition of activities for the new group. Zhao et al. proposed a probabilistic statistical model in 2013, applying a new supervised approach to jointly extract relevant information from videos to capture relationships between objects [3]. It is an undirected model and a three-layer Bayesian probability model.

2.2. Grammar models

Grammar models are used to describe sentence structures and grammar rules in natural language. Their working principle involves defining a grammar structure. They have wide-ranging applications, such as syntactic analysis, semantic analysis, machine translation, and more. Grammar models are crucial for natural language processing tasks as they aid in understanding and generating natural language text by providing information about sentence structure and grammatical correctness. By using appropriate grammar models, sentence structures and grammar rules can be better handled, thereby improving the effectiveness and accuracy of natural language processing.

In 2012, Sofia et al. utilized a grammar-semantic model framework for crowd detection and tracking [4]. This method employed a scene recognition framework named ScReK, which better captures scene information. It also facilitates the modeling of semantic knowledge related to the considered application domain (e.g., people and actions in the scene) and utilizes spatiotemporal constraints to establish correlations with the detected content, thereby enhancing the accuracy of crowd recognition.

Figure 1. Description of the video understanding system [4]

In the framework illustrated in Fig. 1, the video sequence is abstracted into physical objects and utilized for group behavior recognition.

3. Group Recognition Based on Deep Neural Networks

Compared to traditional group recognition methods, using deep neural networks offers significant advantages as it eliminates the need for manual parameter tuning or labeling in traditional approaches. In deep neural network models, two widely used networks are CNNs and RNNs.

CNN is a type of deep learning model primarily applied to image and video processing tasks. Manifold convolutional layers, a few pooling layers, and fully connected layers are the elements of CNN. CNNs have properties like parameter sharing and sparse connections, which have led to their widespread application and excellent performance in the field of group recognition.

RNN is a neural network model with recurrent connections. Unlike traditional feed-forward neural networks, RNNs can handle sequential data such as text, audio, or time-series data that have temporal dependencies. The key feature of RNN is its ability to remember previous inputs through loops, enabling a more comprehensive understanding of contextual information and improving the accuracy of group recognition.

Therefore, by leveraging deep neural networks, specifically CNNs and RNNs, group recognition can be performed more effectively and accurately compared to traditional methods.

3.1. Convolutional Neural Networks

CNNs have been widely applied in the field of group recognition, with numerous specialized architectures. Some examples include Iterative Deep Learning Model (IDLM) [5], multi-Stream CNN [6], Graph Convolutional Network (GCN) [7], and Mask R-CNN [8].

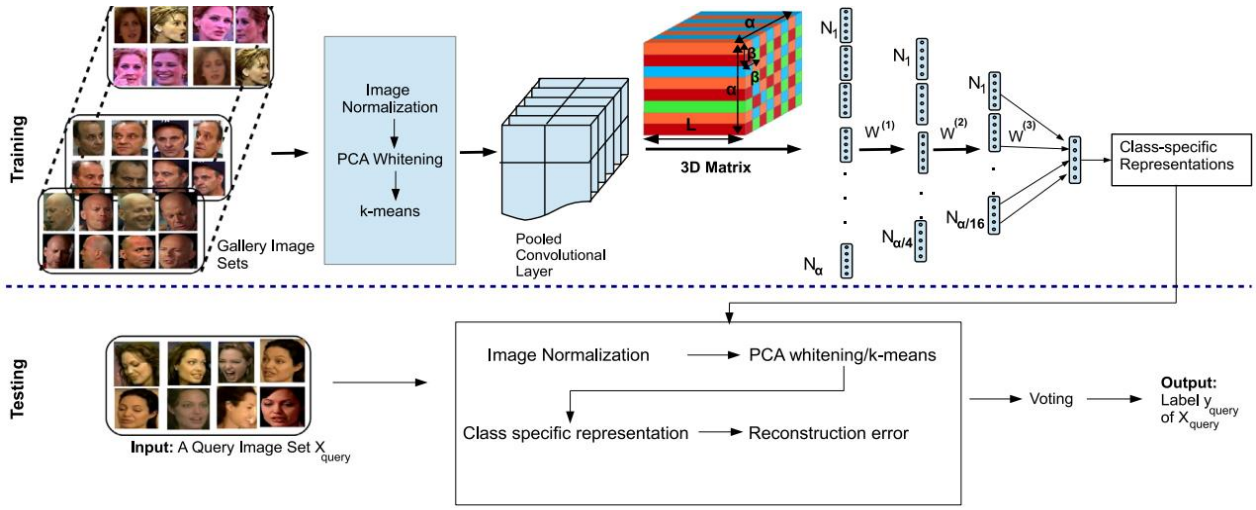

Figure 2. Block diagram of the proposed IDLM for image set-based face/object recognition [5]

The picture shown in Fig. 2 illustrates the framework of the IDLM, which consists of multiple Artificial Neural Networks (ANNs) organized in a hierarchical tree structure for iterative learning of layered feature representation.

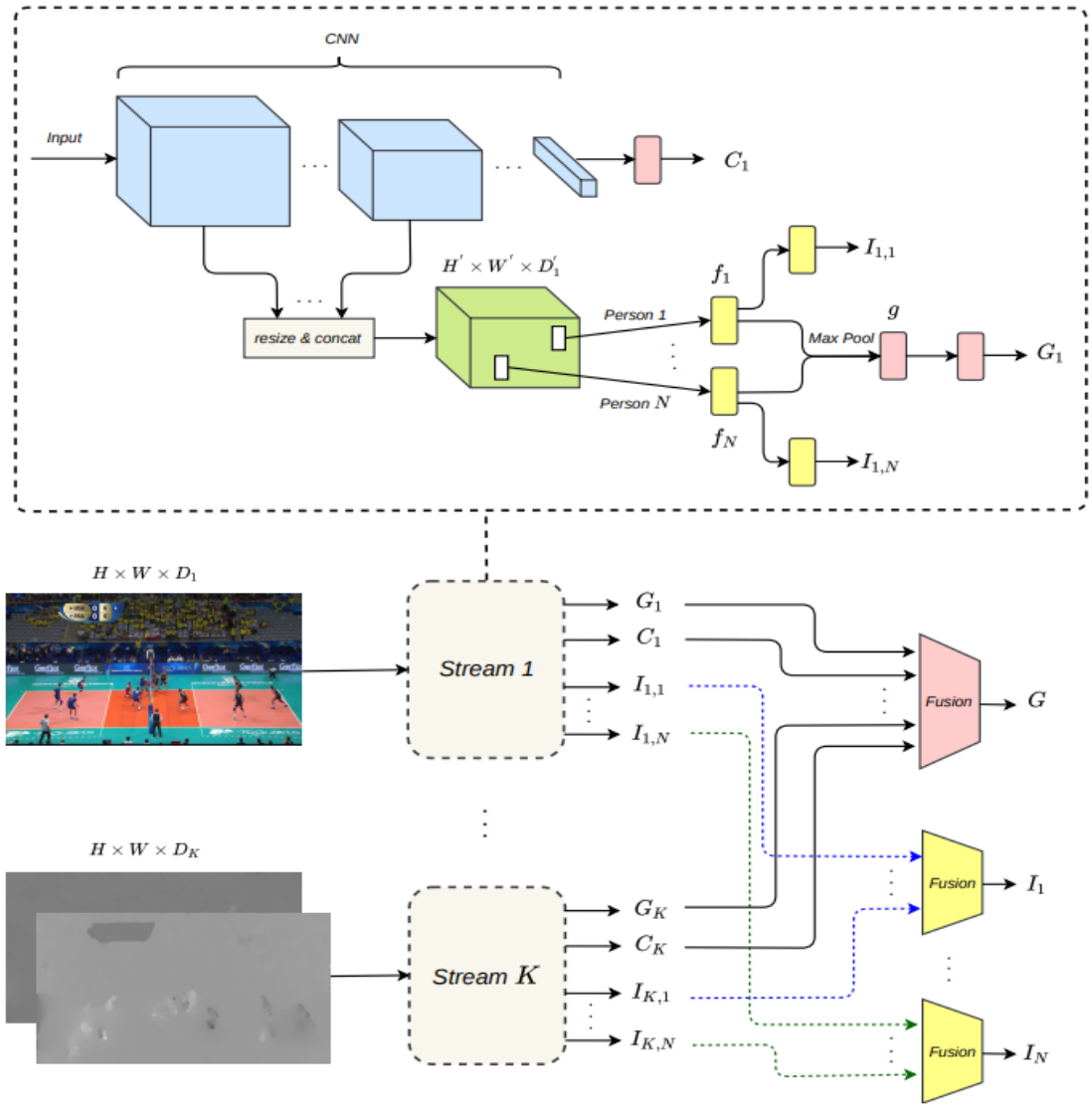

Figure 3. The convolutional framework based on multi-stream [6]

The architecture depicted in Fig. 3 is a multi-stream convolutional architecture. Different streams operate on different modalities and their outputs are getting together to obtain the output. The streams can resize and concatenate features from layers in CNN, which can generate feature maps. These feature maps might extract regions of all the actors in the ground and be used for predicting group activities at both individual and scene levels.

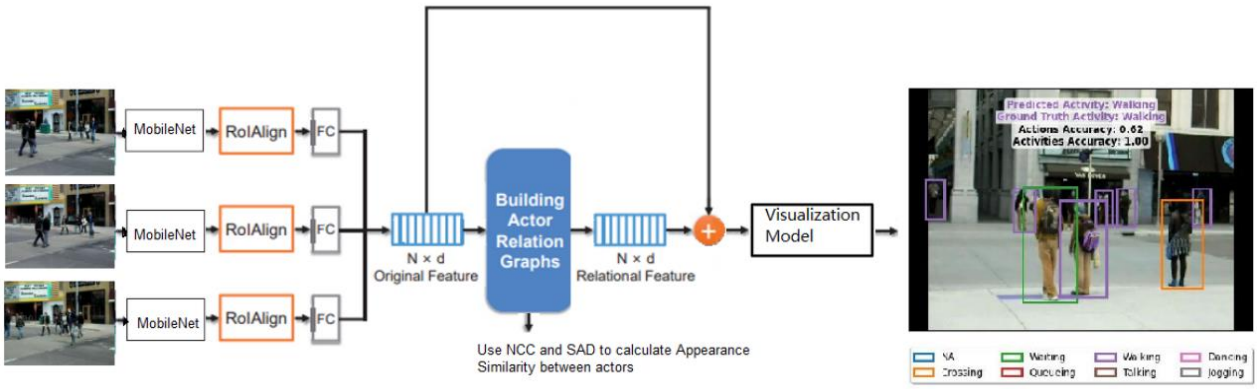

Figure 4. The improved group recognition model with ARG [7]

As shown in Fig. 4, the researchers improved the ARG model by applying a new layer: MobileNet, which is good at picking up feature maps. Normalized Cross-Correlation (NCC) and Sum of Absolute Differences (SAD) are the main methodologies to compute the similarity of appearance in pairs and thus constructed a relational graph among the participants.

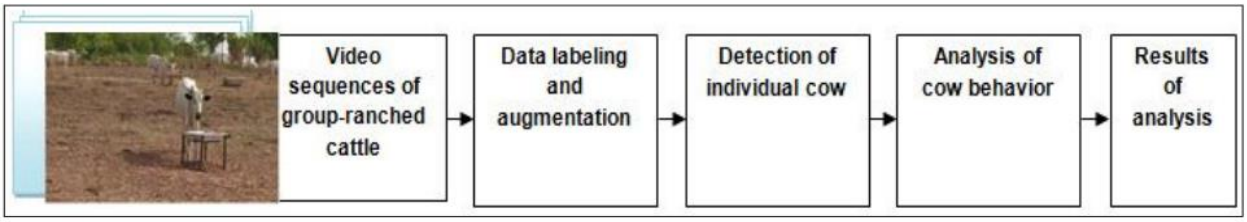

Figure 5. Process-flow of cattle behavior recognition [8]

As shown in Fig. 5, this framework achieves video extraction from cows, data labeling, pre-training, and ultimately behavior recognition and data analysis.

CNN models are particularly suitable for handling image and visual problems. They possess several advantages, including:

1. Translation Invariance: CNNs exhibit robustness to translation, rotation, and scale changes.
2. Parameter Sharing: CNNs significantly reduce the number of trainable parameters by using shared convolutional kernels.
3. Local Perception: CNNs preserve crucial information while reducing redundancy by focusing on local regions.
4. Hierarchical Abstraction: CNNs learn complex semantic features and can perform feature segmentation.
5. Parallel Computing: CNNs enable faster training and testing speeds, leading to improved model efficiency.

Due to these advantages, CNNs are well-suited for group recognition tasks, allowing for accurate and efficient analysis of visual data.

Table 1. Analysis of method applications and characteristics in CNN models

Method	Year	Characteristics	Special Techniques	Function
IDLM [5]	2015	Solves the problem of discriminative information loss in classification tasks	Pooling Convolutional Layer (PCL)	Learns low-level translation-invariant features
			Hierarchical Iterative Artificial Neural Network (ANN)	Learns discriminative nonlinear feature representation of input image sets
Message Passing CNN Model based on Group Activities [9]	2015	① Combines graph models and deep networks for convenient object classification ② Utilizes fine-tuned CNNs and message-passing neural networks	Activity and Interaction Hierarchical Representation	Helps recognize scene information and capture individual activities within a group
			Bag-of-Words Model	Describes local features
Multi-Stream Convolutional Architecture [6]	2018	Integrates multiple streams with CNN, focusing on multi-person behavior recognition in small regions of videos, captures detailed information within small regions	Each stream has two branches	① The first branch focuses on individual action recognition ② The second branch handles group activity recognition in the entire video
GCN [7]	2020	Introduces visualized model: Actor Relation Graph (ARG), combines input video frames, predicted actor's bounding, single and group actions together	NCC	Evaluates the similarity between pairs of image frames
			SAD	Calculates appearance similarity between objects
Visual Semantic Graph Neural Network-Pose, Appearance, Location Attention Learning (VSGNN-PAL) [10]	2022	Uses GCN to accurately capture appearance and motion information while preserving complex scene information	Modal Visual Graph	Encodes the appearance and motion interaction context among multiple individuals
			① Language embedding ② GPU-based Global Reasoning Approach	Refines visual features of individual objects using semantic context and a bidirectional mapping strategy
Performance Evaluation of Mask R-CNN Model [8]	2022	Through four steps: Finally validated that the mask R-CNN model outperforms R-CNN, YOLOv3, and YOLOv4 models in recognition performance	① Extracting video sequences; ② Data labeling and augmentation implementation; ③ Using transfer learning principles, pretraining the model, and selecting the most suitable one; ④ Analyzing individual objects	

From Table 1, the main characteristic of using CNNs can be seen as the utilization of convolutional operations, which focus on the local information of input data. This allows each part of the data to be specialized and maintains local invariance. Additionally, in experiments, combining other methods such as hierarchical iteration [5], multi-stream techniques [6], and language embedding [10] can assist CNNs in accomplishing specific functionalities. These include representing nonlinear inputs, distinguishing between individuals and groups, and enhancing the application of contextual information. This enables better performance in group recognition tasks.

3.2. Recurrent Neural Networks

In addition to CNNs, RNNs also have extensive applications in group recognition. The most commonly used type of RNN is the Long Short-Term Memory (LSTM) model [11, 12]. LSTM was proposed as an improved version of traditional RNNs to address the issues of vanishing or exploding gradients when dealing with long-term dependencies. Compared to traditional RNNs, LSTM introduces gate mechanisms, including input gate, forget gate, and output gate, to control information flow and memory retention. Through these gate mechanisms, LSTM can selectively ignore unimportant information while retaining and updating important information, thus effectively capturing long-term dependencies in sequential data. This makes LSTM perform remarkably well in tasks such as natural language processing, text generation, and speech recognition.

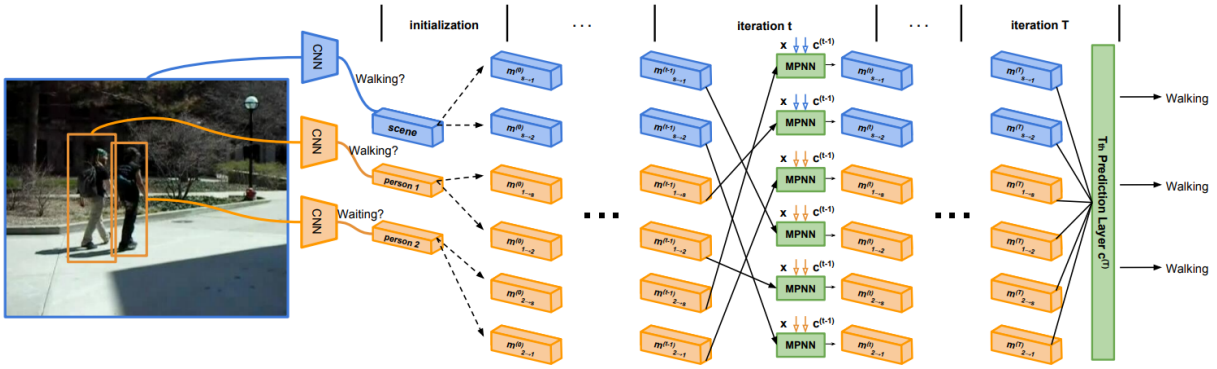

Figure 6. The pipeline of inference in an RNN [13]

As shown in Fig. 6, this model utilizes unary scores for information initialization. In each iteration, relevant message units, unary scores, and previous output predictions are used to compute new information.

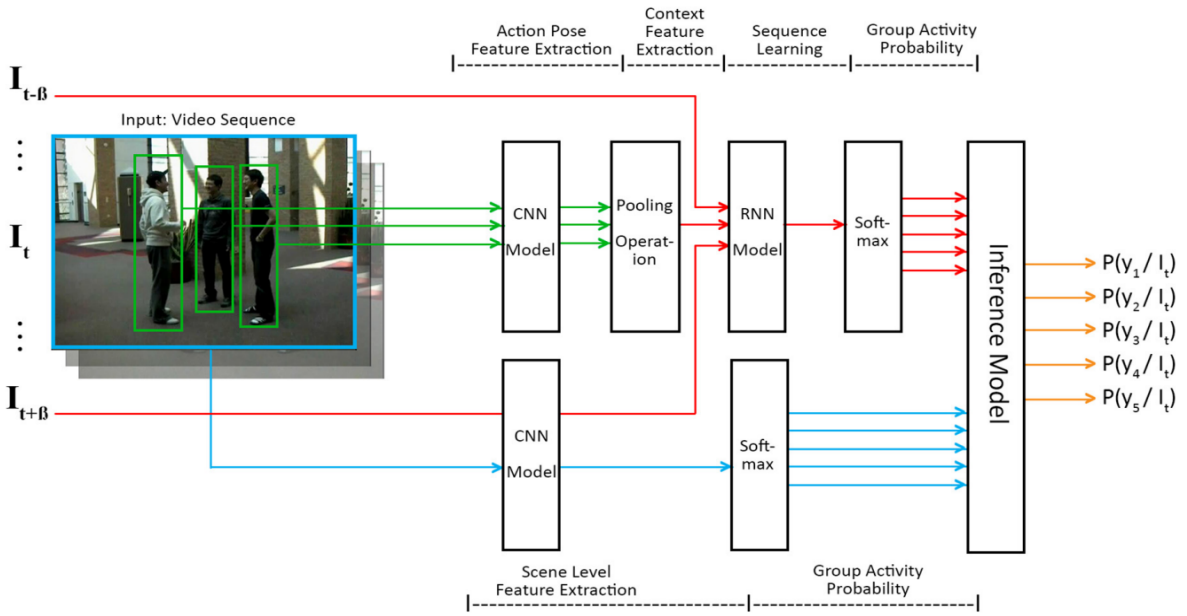

Figure 7. Contextual deep neural network architecture [12]

As shown in Fig. 7, this model first employs pre-processing to obtain information about individuals in the video sequence. Then, an RNN model is used for learning, and the obtained contextual features are trained using a softmax classifier for recognition, ultimately completing the recognition process.

Table 2. Application and Feature Analysis of Methods in RNN Models

Method	Year	Characteristics	Special Techniques	Function
Model for LSTM [11]	2016	Combines LSTM model with a hierarchical model to represent individual's action dynamics and aggregate personal-level information	Two different LSTM models	Capture individual actions and relationships between individuals
			Hierarchical deep model	Aggregate features and relationships among participants
Gated Recurrent Unit (GRU) [13]	2016	Combines RNN with gate functions to detect relationships between participants and record context information	Gate function approach	Used in learning structures for training purposes, facilitates recognition of other models
Model for LSTM [12]	2019	Designs a model combining CNN and RNN at the human action-pose level	① multi-layer architecture ② Contextual relationship	approach for comprehensive temporal analysis of group activities
			Gated Recursive Unit	Aggregates recognized human actions and poses for subsequent recognition applications

From Table 2, the key feature of RNNs is their ability to preserve and utilize current information for comprehensive information processing in the next time step, enabling better attention to contextual information. In experiments, special techniques have also been applied, such as the use of gate methods [13], which can identify learning actions and be used for group recognition tasks.

3.3. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) Fusion

Compared to the previous two approaches that solely employ either CNN or RNN for group recognition, combining CNN and RNN allows both models to leverage their respective strengths. This fusion enables the processing of objects that possess both temporal and spatial characteristics, thereby improving recognition performance. For example, in video analysis, this approach retains both temporal and spatial information, leading to more accurate and efficient recognition.

In 2018, Rudy et al. applied an improved RNN matching strategy by combining RNN and CNN for remote learning [14]. They utilized unsupervised dictionary learning methods to capture input features and enhance the accuracy and speed of person recognition. Building upon dictionary learning, they introduced fully connected layers with rectified linear unit (ReLU) activation and drop out, creating a novel Siamese network. This network minimizes the size of recognition frames, making it easier to distinguish different objects.

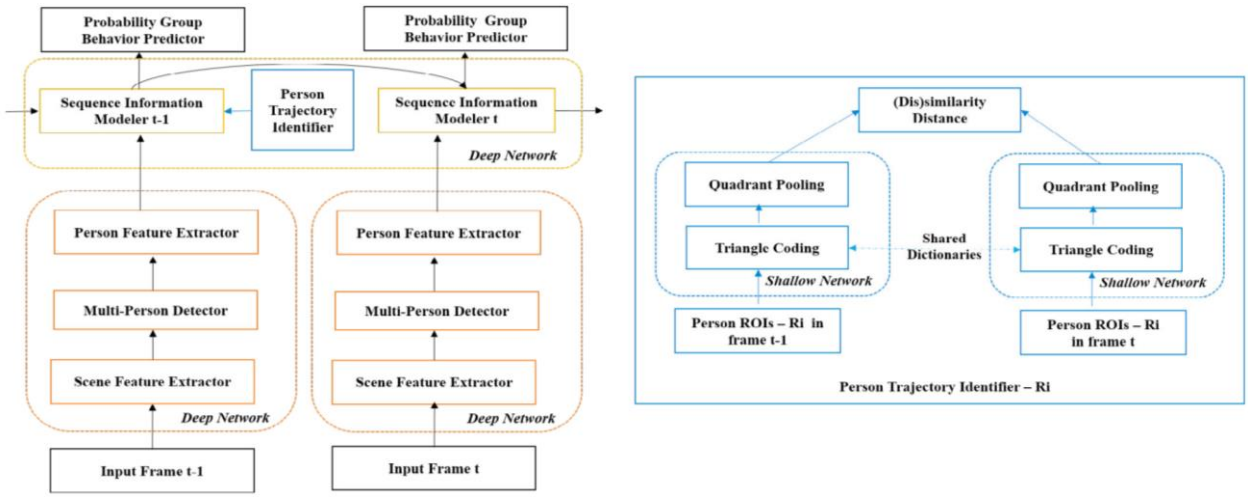

Figure 8. The group behavior recognition architecture [14]

As shown in Fig. 8, this experiment combines RNN and CNN using dictionary learning methods to accomplish group behavior recognition. It involves techniques such as scene feature representation, multi-person detection, task feature representation, sequence information modeling, person trajectory recognition, and probabilistic group behavior prediction. These methods collectively enable the recognition of group behavior.

3.4. Pose Analysis

In group behavior recognition, another major category is posing analysis. This method typically represents various objects identified in a video as a combination of lines and nodes. The human form is represented by lines representing the torso and limbs, with nodes representing major joints. This simplifies the objects while preserving the most fundamental features to capture the temporal and spatial characteristics of the objects. It often utilizes manifold learning algorithms to map high-dimensional skeletal data to low-dimensional manifold space.

Table 3. Application and Characteristics Analysis of Methods in Pose Analysis Models

Method	Year	Characteristics	Special Techniques	Function
Temporal-spatial structure modeling of human body posture [15]	2013	Improved a method for estimating the positions of human body joints, capturing the spatial configuration (spatial subset) of body parts in a framework and the body part motion (temporal subset) of human motion characteristics. These subsets are obtained using an efficient contrast-minimizing algorithm	Spatial domain, data mining techniques	Obtain unique spatial configurations (postures) that co-occur with body parts
			Temporal domain	Obtain a collection of unique co-occurring posture sequences of body parts
			Based on posture action recognition	Merge subset segmentation of all frames and temporal constraints to infer the best posture, providing more accurate localization of body joints
Manifold-based skeleton action recognition [16]	2017	Establish a rotation mapping layer, rotation pooling layer, and logarithmic mapping layer to optimize the input object features, reduce high-dimensional features, and classify output information reasonably, achieving ideal recognition effects	Manifold learning algorithm	Map high-dimensional skeleton data to low-dimensional manifold space
			① Backpropagation framework ② Variant of stochastic gradient descent	Optimized deep models, used classical SGD algorithm, solved gradients using the chain rule, in the rotation layer, used loss function method to compute required Riemannian gradients
Recognition framework based on spatiotemporal features of posture [17]	2020	① A new framework that uses 3D skeleton information to recognize human behaviors. The main components are posture representation and encoding ② Using Fisher vectors (FV) to encode and using extreme learning machine (ELM) to train to get the descriptors	Pose descriptors composed of three elements	① Normalized position of the joint ② Time displacement information 1) predefined 2) timestamp
			Principal Component Analysis (PCA)	Avoid high-dimensional problems of descriptors
Action X Pose: A novel posture-based 2D pose-level HAR method [18]	2020	Action X Pose picks up different level characteristics from posture and provides them to LSTM neural networks and one-dimensional CNNs for classification. It reaches the top of accuracy and has great robustness to occlusion and data-loss	Human Activity Recognition (HAR)	Techniques and methods for recognizing and classifying human activities
			New posture-based HAR dataset ISLD	Recorded real posture-level HAR in an intelligent sensing laboratory in a closed-circuit television-like environment
			Extensively explore data augmentation which can make the proposed model better with nowadays datasets	
Dual-stream neural network [19]	2021	One stream is a self-attention GCN (SAGCN) that picks up temporal and spatial information, and the other stream is a residual connection-enhanced bi-directional independent RNN (R Bi-Ind RNN) that gets long-term temporal information	SAGCN	Besides the stable topological structure and feature in GCN, it also has a dynamic SA mechanism that adaptively utilizes the relationships between all hand joints
			Residual connection-enhanced Bi-Ind RNN	Extended Ind RNN with bi-directional processing capability for temporal modeling

From Table 3, what can observe is the characteristics and functions of various pose analysis models. Pose analysis represents the human body using lines, enabling better identification of individuals and ensuring recognition accuracy. In the experiments, several methods were employed to assist pose recognition, including spatial and temporal domain approaches, self-attention mechanisms, etc. These methods facilitate spatial information configuration, temporal information integration, and adaptive

utilization of all available information, ultimately aiding in recognition and improving recognition accuracy.

4. Comparison of Datasets and Algorithm Performance

4.1. Datasets

In the field of crowd recognition research, a variety of datasets have been used, including general-purpose datasets and specialized datasets. Here are some commonly used datasets.

During the early stages of traditional algorithm development, researchers had limited options and sometimes had to collect their own videos and construct the required datasets. Therefore, only a brief introduction will be provided here. For example, in 2009, Wei et al. collected some sports videos from the internet and created a dataset for collective activities, which included four types of sports: badminton, tennis, football, and basketball [1]. Over time, more datasets emerged. In 2013, researchers used the behave dataset [2], which is a commonly used dataset for behavior analysis and recognition. Each behavior sequence in this dataset is labeled with different categories such as "walking," "running," "talking on the phone," etc. Another example is the IDIAP dataset [2], which is widely used in speech recognition and facial recognition domains and was created and maintained by the IDIAP Research Institute in Switzerland. The IDIAP dataset includes multiple sub-datasets, with some of the most famous and commonly used ones being the TIMIT benchmark dataset for English speech recognition, the BANCA dataset for multimodal face and speech biometric research, and the MOBIO dataset for multimodal biometric recognition.

Table 4. Introduction of commonly used datasets in crowd recognition

Dataset	Video Count	Clip Count	Recognized Behaviors	Behavior Categories	Training Set	Test Set
NUS-HGA		476	Everyday multi-person behaviors	6		
BEHAVE		174	Everyday multi-person behaviors	10		
UCF-Crime		1900	Criminal behaviors	14	1610	290
Collective Activity	44	2500	Different types of group activities	① 6 types of individual behaviors ② 5 types of group behaviors	Two-thirds of the total count	One-third of the total count
Volleyball	55	4830	Actions in volleyball match videos	① 9 types of individual behaviors ② 8 types of group behaviors	3493	1337
NCAA Basketball	257	6553	Actions in basketball matches	11		
NBA Dataset	181	9172	Actions in basketball matches	9	83%	17%
Collective Sports	257	2187		① 5 types of individual behaviors ② 11 types of group behaviors	80%	20%
UT Interaction	20		Interactions between individuals	6		
UCF101		13320	Various human behaviors	101		

In the deep learning stage, with the rapid development of the Internet, various datasets have emerged, greatly facilitating video-based crowd recognition. In Table 4, what can be observed is that deep neural networks generally use larger datasets compared to non-deep learning methods, as seen

in the first two datasets listed in Table 4. The size of these datasets is typically 10-100 times larger than the others. Consequently, the most notable difference is that deep learning methods, due to their exposure to more data during training, can better recognize complex scenarios. Among these datasets, the most commonly applied ones are the collective activity dataset and the volleyball dataset.

The Collective Activity Dataset consists of videos and images that simulate various collective behavior scenarios. Researchers can use this dataset to explore and address issues related to collective behavior, such as crowd management, traffic flow analysis, and social interaction research. It serves as a foundation and reference for algorithm development, training, and evaluation. These datasets play a vital role in enhancing the understanding of group behavior and driving research in related fields.

The Volleyball Dataset is commonly used for tasks such as action recognition, detection, and generation in the field of deep learning. It comprises videos of indoor volleyball matches with annotations for on-court actions. The dataset includes video clips from different scenarios and action categories, such as serving, spiking, passing, and blocking. Each video clip is associated with action labels and can be used for training and evaluating deep learning models, primarily for action recognition and action detection purposes.

4.2. Performance Comparison of Traditional Models

Table 5. Analysis of Results for Traditional Model Approaches

Classification	Reference	Method	Year	Dataset	Method Comparison	Average Accuracy (%)
Probabilistic Graph Models	[1]	Probabilistic Summation	2009	Constructed Collective Activity Dataset, including four sports: Badminton, Tennis, Soccer, Basketball	①Template Matching ②Probabilistic Summation	85 88.48
	[2]	Probabilistic Framework	2013	BEHAVE IDIAP	Gaussian Process Model	93.65 92.3
	[3]	Probabilistic Statistics	2013	Unstructured Social Activity Attribute (USAA) dataset	①Med LDA ②g-Class-RBM ③Relational Topic Model (RTM)	40.57 44.35 50.33
Grammar Models	[4]	Semantic Grammar Framework (ScReK)	2012	①Turin Metro Surveillance ②Eindhoven Airport Surveillance ③CAVIAR3	①Characteristics-based Person Detection ②Semantic Grammar Framework	73 95-98

From Table 5, it can be seen that the recognition accuracy of traditional models can indeed improve compared to basic models. However, the accuracy fluctuates greatly, and different methods show significant differences on different datasets. Some methods perform well with accuracy above 90%, while others only have around 50% accuracy, indicating that the performance of traditional models is not particularly good.

After incorporating relevant information from the literature, some possible reasons are provided for the average accuracy of only 50%. The literature used sparse Bayesian learning methods, which may lead to low recognition accuracy due to factors such as insufficient data and strong feature correlation. Although the dataset USAA used in the literature consists of 1466 videos covering 8 categories, the dimensionality is high at 5000 dimensions. The high dimensionality could contribute to the low recognition accuracy. Additionally, since it involves group recognition, the strong correlation between objects can also affect the recognition results. In conclusion, these factors contribute to the relatively low accuracy of 50%. However, it should be acknowledged that compared to other literature, the model presented in this paper still improves recognition accuracy.

4.3. Comparison of Deep Network Performance

Table 6. Analysis of results from deep model methods

Classification	Reference	Year	Dataset		Method Comparison	Average Accuracy (%)
CNN	[5]	2015	Face Recognition	① YouTube Celebrities (YTC) ② Honda/UCSD ③ CMU Mobo	IDLM	76.52 100 98.41
			Object Classification	ETH-80		98.64
	[9]	2015	Collective Activity Dataset		① Graph Model Spatial Information ② Hierarchical Graphical Network	79.1 80.6
			Nursing Home Dataset		① Pure Deep Learning ② Hierarchical Graphical Network	79.15 80.2
	[6]	2018	Volleyball Dataset		① Individual ② Group	82.36 88.56
			Collective Activity Dataset		Multi-stream Model (MCA)	95.17
	[7]	2020	Collective Activity Dataset and Augmented Activity Dataset		① Inception-v3 ② Mobile Net	91.81 90.41
					① Embedded Dot Product ② NCC ③ SAD	92.71 93.50 93.98
	[10]	2022	Cattle Action Dataset		Mask R-CNN Model	92.065
RNN	[11]	2016	Group Activity		① Context Model ② Deep Structural Model ③ LSTM Model	79.1 80.6 81.5
			Volleyball Dataset		LSTM Model	51.1
	[13]	2016	① Collective Dataset (Extended) ② Nursing Home Dataset		RNN Network with Gate Functions	81.2 (90.23) 85.5
	[12]	2019	Collective Activity Dataset		① LSTM Model ② GRU	82.94 83.45
Combination of CNN and RNN	[14]	2018	① Volleyball Dataset ② CUHK03		Combination of CNN and RNN with Unsupervised Dictionary Learning	91.4 98.86
Pose Analysis	[15]	2013	① UCF Sports Dataset ② MSR-Action3D Dataset ③ Keck Gesture Dataset		Method for Estimating Human Joint Positions	90.00 90.22 97.62
	[16]	2017	① G3D-Gamingdataset ② HDM05dataset ③ NTU RGB + D dataset		Lie Group Model	89.1 75.78 66.95
	[17]	2020	① UT Kinect Action Dataset ② Florence3D Action Dataset ③ UTD-MHAD Dataset		3D Skeleton Information Framework	97 89.6 92.6
					Supervised Learning Algorithm (SVM)	89 24 86.04
	[18]	2020	Pose-based HAR Dataset ISLD		Action X Pose: A Novel Pose-level HAR Method Based on Pose	95.11
	[19]	2021	① Dynamic Hand Gesture Dataset DHG ② First-Person Hand Action Dataset FPHA		Deep Residual Bidirectional Ind RNN (R Bi-Ind RNN)	95.18 90.26

In Table 6, it is evident that when using deep networks, the accuracy of group recognition is relatively high. Although there are exceptions with 51.1% and 67% [11, 16], the majority of accuracies are above 80% and relatively stable. In comparison to traditional methods, although some traditional methods achieve even 100% accuracy, their accuracy fluctuates greatly and is not very stable. Additionally, due to the large amount of data in deep network datasets, the accuracy is not as good as traditional methods for some smaller datasets.

Furthermore, the advantage of deep methods can also be seen in processing speed. For example, in Method 4, a graph convolutional network is used [7], and it is about 2500 seconds faster than traditional Inception-v3 on the same dataset, reducing processing time by one-fourth. This demonstrates the superiority of deep networks over traditional networks in terms of speed.

Overall, group recognition in deep networks is superior to traditional methods and more stable. However, the contributions made by traditional methods should not be disregarded. Both methods are valuable approaches.

5. Conclusion

Group recognition, as a popular and important topic in the field of computer vision, has received widespread attention in today's society. This paper reviewed group recognition from the perspectives of traditional models and deep models based on the current research status in the field. This paper also analyzed the characteristics and applications of commonly used datasets. Furthermore, some analyses are provided on the combination of old and new methods, improving accuracy through multiple model combinations, and recognition in special scenarios.

With the development of deep learning and the maturity of technology, significant changes are happening in this field. Although the current recognition technology has reached a relatively high level, there is still room for improvement to meet human expectations. Based on existing research and the latest research directions, this paper will provide prospects for deep learning-based group recognition.

1) Incompatibility of combining old and new methods. The LSTM method is already a mature approach with extensive applications in video recognition. The combination of CNN-RNN models has emerged in recent years and has demonstrated higher recognition accuracy than older methods. However, when combining LSTM and CNN-RNN models, a problem of lower recognition accuracy arises, confusing recognition. Therefore, in future research, better methods can be explored to address the compatibility issues between old and new methods.

2) Combination of multiple models. In recent years, scholars have attempted to combine CNN and RNN models, achieving promising results. This indicates that it is possible to combine various effective models to complement each other's strengths, leading to a better overall model. Therefore, in the future, different combinations of models can be explored, which may yield unexpected results.

3) Recognition in special scenarios. Research has found that for the recognition of certain special objects, such as the behavior of cows, there can be a lot of interference information, resulting in misclassifying non-recognizable objects as the desired recognizable objects. This significantly increases the error rate of recognition. Therefore, recognition in such special scenarios remains a highly potential direction for development.

References

[1] Wei, Qingdi, et al. Group Action Recognition Using Space-Time Interest Points. State Key Laboratory of Pattern Recognition, 2009, 1 - 11.

[2] Yin, Yafeng, et al. Small group human activity recognition. IEEE International Conference on Image Processing, 2013, 1 - 13.

[3] Zhao, Fang, et al. Relevance Topic Model for Unstructured Social Group Activity Recognition. 2013, 2580 - 2588.

[4] Zaidenberg, Sofia, B. Boulay, and F. Bremond. A Generic Framework for Video Understanding Applied to Group Behavior Recognition. IEEE 2012.

[5] Shah, Syed Afaq Ali, M. Bennamoun, F. Boussaid. Iterative deep learning for image set based face and object recognition. Neurocomputing 174. JAN. 22PT.B, 2016: 866 - 874.

[6] Azar, Sina Mokhtar Zadeh, M. G. Atigh, A. Nickabadi. A Multi-Stream Convolutional Neural Network Framework for Group Activity Recognition, 2018.

[7] Kuang, Zijian, and X. Tie. Improved Actor Relation Graph based Group Activity Recognition, 2020.

[8] Bellot, Rotimi Williams, et al. Behavior Recognition of Group-ranched Cattle from Video Sequences using Deep Learning. Indian Journal of Animal Research, 2022 4: 56.

[9] Deng, Zhiwei, et al. Deep Structured Models for Group Activity Recognition. arXiv 2015.

[10] Liu, Tianshan, et al. Visual-semantic graph neural network with pose-position attentive learning for group activity recognition. Neurocomputing 2022, 491: 217 - 231.

[11] Ibrahim, Moustafa, et al. A Hierarchical Deep Temporal Model for Group Activity Recognition., 10.48550/arXiv.1511.06040. 2015.

[12] A, S. A. Vahora, and N. C. C. B. Deep neural network model for group activity recognition using contextual relationship. Engineering Science and Technology, an International Journal 2019, 22. 1:47 - 54.

[13] Deng, Zhiwei, et al. Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition. Neurocomputing, 2016.

[14] A, Rudy Cahyadi H. P, J. F. B, and H. K. P. A. Group Behavior Recognition Based on Dictionary and Hierarchical Learning. Procedia Computer Science 2018, 144: 52 - 59.

[15] Wang, Chunyu, Yizhou Wang, and Alan L. Yuille. An approach to pose-based action recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2013.

[16] Huang, Zhiwu, et al. Deep Learning on Lie Groups for Skeleton-based Action Recognition. IEEE Computer Society, 2016.

[17] Agahian, Saeid, Farhood Negin, and Cemal Köse. An efficient human action recognition framework with pose-based spatiotemporal features. Engineering Science and Technology, an International Journal 2020, 23.1: 196 - 203.

[18] Angelini, Federico, et al. 2D pose-based real-time human action recognition with occlusion-handling. IEEE Transactions on Multimedia 2019, 22.6: 1433 - 1446.

[19] Li, Chuankun, et al. A two-stream neural network for pose-based hand gesture recognition. IEEE Transactions on Cognitive and Developmental Systems 2021, 14.4: 1594 - 1603.