IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 7, July 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207051 |

Optimized Aerial Activity Recognition via Integrated GAN and CNN-LSTM Fusion Techniques

Manjunath A[1], Rajeshwari N [2]

MCA Student, Department of Computer Application, Bangalore Institute of Technology, Bangalore, India[1]

Assistant Professor, Department of Computer Application, Bangalore Institute of Technology, Bangalore, India[2]

ABSTRACT: Unmanned aerial Vehicles also known as drones are applied widely**.** Drones can identify human activities in videos but face challenges due to camera issues and limited data. This paper proposes a method combining data manipulation and a special technique to create more training data. This technique produced very accurate results using real-world videos.

KEY WORDS: Aerial Vehicles, Special technique

I. INTRODUCTION

Drones have transitioned from primarily military use to various civilian applications in smart cities. They offer advantages like being autonomous, fast, and cost-effective. Drones are employed in video surveillance, research, security, rescue, and disaster management. They can improve efficiency, security, and general well-being in urban environments, and They have a lot of potential for transforming smart cities.

It can take a long time and be expensive to gather and categorize the large amount of training information required for deep being an able to recognize human activity in drone videos. Unbalanced data might also reduce the effectiveness of models. Researchers use strategies for data augmentation, such as more recent approaches like GANs and more conventional methods, to get over these obstacles.

This paper suggests a hybrid strategy to produce helpful features for the classification of actions in aerial videos by combining data transformation and GAN-based techniques. This method generates precise sets of features that capture the attributes of various actions by conditioning GANs on particular class levels. Accurate action recognition and classification can be achieved by training SoftMax classifiers with the CNN-LSTM functionalities that were created. This technique overcomes the shortcomings Some video examples produced using GANs while lowering computing complexity. The study is organized into sections that provide an overview of relevant literature, the suggested methodology, findings from the experiment, and conclusions.

II. RELATED WORKS

Spotting human actions in videos (HAR) is a hot topic in computer vision. Techniques have come a long way, eschewing manually created features to deep learning that automatically learns what to look for, making HAR much more powerful.

2.1 Handcrafted Methods

Previous research on human action recognition (HAR) relied on manually created characteristics that are determined by video footage. Three groups can be formed from these techniques:

1. Body model-based approaches: extract outline of the human body for computation skeleton or reconstruct 3D pose.
2. Holistic approaches: depict overall body dynamics using silhouette or optical flow/gradient-based methods.
3. Local feature-based approaches: use Interest Points (IPs) and bag-of-words (BOW) representation to efficiently represent actions.

However, handcrafted methods have limitations, such as:
- Relying on specific assumptions

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207051 |

- Insufficient generalization and resilience capabilities mean that a model or system doesn't do well in unfamiliar situations., unseen data or in different conditions than it was trained on. It struggles to adapt and maintain accuracy outside of its initial training environment.

These limitations can be defeated with deep learning methods, which can learn features automatically and improve performance.

2.2 Deep Learning techniques

In human action recognition (HAR), deep learning techniques—in particular, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)—have demonstrated effectiveness. RNNs and Networks using Long Short-Term Memory (LSTM) are better when obtaining temporal characteristics than CNNs are at extracting spatial features. However, CNNs struggle with temporal changes in human behavior, and RNNs face challenges with long sequences. Combining CNNs and LSTMs has shown promise, but limited aerial datasets have hindered their application in identification of airborne action.

This work tackles the problem of having too little and unbalanced data by proposing to use CNN-LSTM models along with the augmentation of data strategies. The CNN-LSTM model can be trained more successfully to identify actions in aerial movies by increasing the quantity and variety of data.

2.3 Data Augmentation

Enhancing HAR systems through the augmentation of data is essential, particularly when dealing with tiny datasets. Simple transformations and sophisticated deep learning techniques are the two basic approaches for doing this.

Basic methods such as combining photos, cutting portions of films, and altering features are examples of simple transformations. Nevertheless, more advanced methods, like the use of Generative Adversarial Networks (GANs), can produce artificial information that closely resembles the original. Creating synthetic films, transforming photos into other styles, and growing the dataset are all made possible by GANs. GANs, however, can struggle with extremely particular data or uneven quality.

We focus on improving HAR systems in this work by using both simple transformations and features generated by GANs. The AugLy package is used for simple transformations, while GANs are used to create synthetic data. Combining these techniques results in training data that is more varied and realistic, which enhances system performance.

III. SUGGESTED METHOD

1. Basic Transformations: We employ easy methods to make the most out of our small dataset.
2. Data Preprocessing: We clean and prepare the unprocessed data for analysis.
3. Feature Extraction: We use CNN-LSTM and 2D CNN to analyse every frame additionally pull-out important details.
4. Motion Processing: We use LSTM to understand the movement patterns and long-term relationships in the data.
5. Synthetic Data Generation: We use WGAN to create extra features, making our dataset bigger. Finally, we classify actions using an activation of SoftMax function.

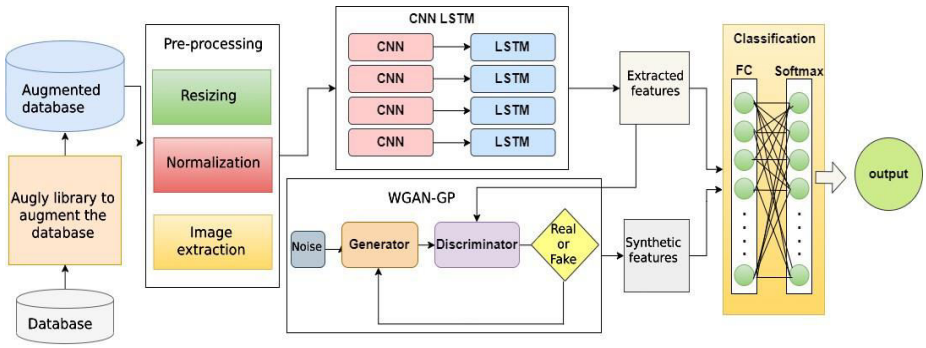

Figure 1: The suggested method for identifying motions made by people in films taken from above.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207051 |

3.1. Phase 1: Data Enrichment

Techniques to enhance data are used to make the dataset bigger and stop the model from overfitting. The AugLy library, which has many ways to change both space and time in videos, is perfect for testing how strong a model is. We chose four techniques to deal with the complexities of aerial videos: adding different backgrounds, blurring, changing brightness as well as increasing Nois. These techniques assist the model navigate obstructions, shifting illumination, and other difficulties in aerial footage by simulating changes in the real world. We are able to exercise more dependable model by intentionally growing the dataset.

3.2. Phase 2: Initial processing

Three procedures were taken to able to pre-process aerial films in relation to human activities recognition:
i) Removing frames from videos t2o0 compile a list of distinct pictures.
ii) putting 20 as the sequence duration and resizing the components of a specified height and width(224x224).
iii) Setting values of pixels ranging from 0 to 1 to a normal
By preparing the information for additional processing, these actions improved the suggested model's efficiency. We can increase the precision of the model by properly structuring the data.

3.3. Phase 3: Highlight Deletion

Our system uses a combination of two categories of neural networks: CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory) to analyse footage. The CNN looks at every in the video's frame to understand the spatial features (like shapes and objects), while the LSTM helps understand the temporal features (how things change over time). Figure 2 shows In our scenario, how these two networks cooperate.

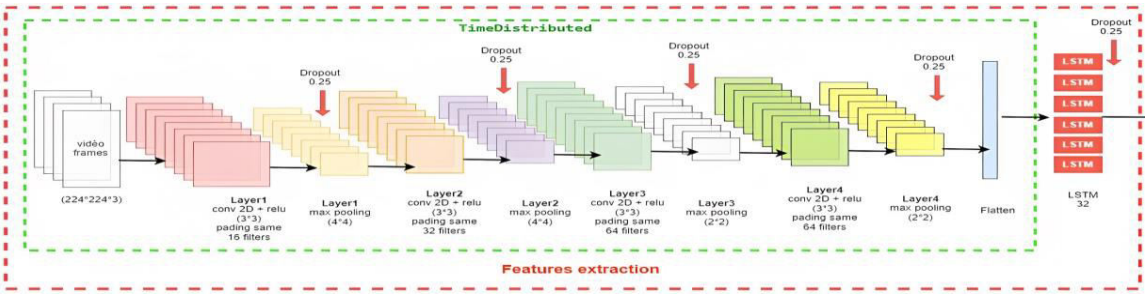

Figure 2. CNN-LSTM architecture model.

3.3.1 Extraction of Spatial Features

We use a neural network type called a 2D-CNN (2D Convolutional Neural Network) to analyze action videos taken from above. The 2D-CNN looks at 20 frames from every video to identify spatial features like, corners, edges and textures. It does this by applying filters to the frames. Because of hardware limitations, we designed our network with four layers for detecting features, four layers for reducing the size of the data, and an output layer that classifies the features.

In simple terms, CNNs are great at finding details in individual frames, while LSTMs (Long Short-Term Memory networks) are better at understanding how things change over time. Our combined model, which blends the two CNNs and LSTMs, effectively captures both the spatial details and temporal changes in video data. Figure 2 displays how the model is organized.

3.3.2 Temporal Feature Extraction

While CNNs are great at extracting details from single frames, they aren't as good at understanding how things change over time. To handle this, we use One kind of RNN intended for sequences are Long Short-Term Memory (LSTM) networks. LSTMs are more skilful than average RNNs in resolving problems that crop up during long sequence processing.

An LSTM has three gates—input, forget, and output—that help it manage information over time.
In our system, we use 32 units in an LSTM are displayed in Figure 3. We feed the spatial characteristics retrieved by the 2D-CNN into the LSTM to understand how these features change over time. By combining CNNs and LSTMs, the video data's temporal and spatial components can both be captured by our approach.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207051 |

The difficulty of our hybrid approach is determined by the sum of the complexity of the CNN and LSTM layers. The following formula can be used to calculate our model's complexity.: O (d × nl × nl-1 × sl × ml + w × i × e), where:
- The convolutional layer count is d1,
- The number of kernels is nl,
- nl-1 represents the quantity of input channels,
- sl the kernel's size,
- ml is the output characteristic map's size,
- I represent the input length;
- w is the quant of LSTM weights;
- where i represents input length.
- One factor connected towards LSTM units is e.

In simpler terms, the formula helps us understand how complex our model is according to the quantity of components.

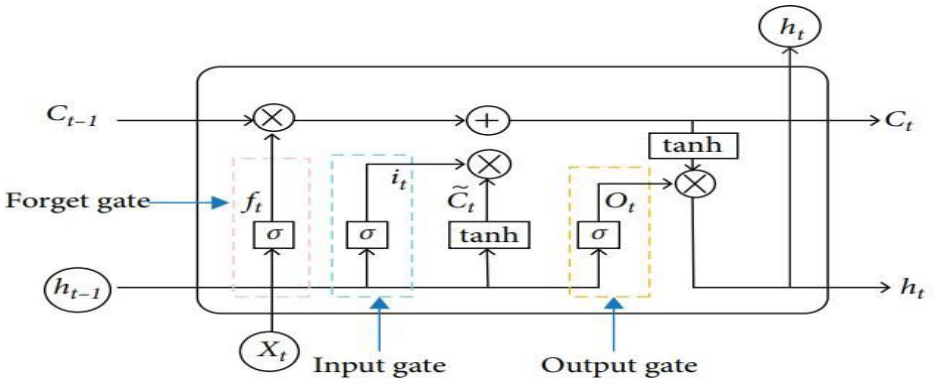

Figure 3. Diagram showing how a long-term memory cell works

Phase 4: WGAN for Producing Artificial Features
We use a model called WGAN-GP to create artificial features based on the spatial and temporal features we extracted earlier. The WGAN-GP model has two parts: a generator and a discriminator.

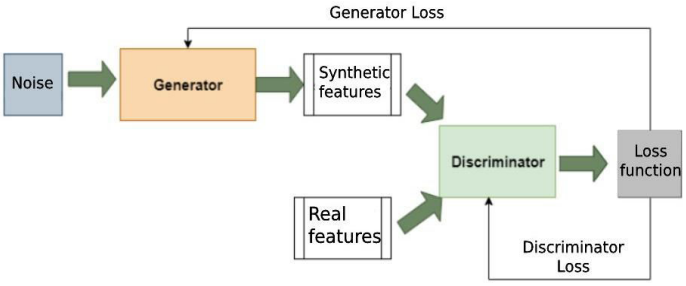

Figure 4. An example of the WGAN-GP model.

The generator in our system starts with random noise and a class label to produce artificial features. The discriminator then compares both artificial and genuine features along with a class label to see how different they are.

We use WGAN-GP because it works better than the original GAN, being more stable and faster to train. WGAN-GP uses a gradient penalty to keep the discriminator in check, preventing unstable training and helping the model learn more quickly and reliably.

Phase 5: Action Classification
Actions are classified using CNN-LSTM and WGAN-GP features. The dense layer's output size matches the number of classes. SoftMax activation determines the highest-likelihood class.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207051 |

IV. EXPERIMENTS AND RESULTS

We use 400 clips from 8 different actions in the YouTube 7 Aerial dataset to test our human action detection algorithm. We analyse the results and compare them with state-of-the-art techniques.

Dataset

For the examination of aerial videos, the dataset for YouTube Aerial is an invaluable tool, with 50 videos per action, each with 320x240 pixels and FPS of 25 are used.

Implementation Setup

We conduct tests with a V100 Tesla and Google Colab-Pro+ and 52 GB RAM. We use Python 3, Jupyter Notebook, Tensorflow, and Keras. We extract 20 frames per video, resize inputs to 224x224x3, and fundamental methods for augmenting data. We divide the dataset (80%, 20%, and 20%, respectively) within sets for testing, validation, and training.

Table 1. CNN-LSTM model hyperparameters that are suggested.

Parameter	Values
frame resizing input	224 ×224 ×3
quantity of CNN tiers	4
Filter sizes	16, 32, 64, and 128
Amount of the Kernel	3×3
Max pooling	Yes
The quantity of LSTM units	32
Epochs	100
Quantity in batch	30
Enhancer	Adam
Loss function	loss of Categorical cross-entropy
Dropout rate	0.25

Table 2. WGAN generator hyperparameters.

Parameter	Values
Dimension z of the noise vector	100
The quantity of the layers	4 layers
Function being activated	LeakyReLU; output layer: hyperbolic tangent
The quantity of each layer's neurons	(128, 256, 512, 32)

Table 3. The discriminator of WGAN-GP hyperparameters.

Parameter	Values
The quantity of layers	4 layers
Functions of activation	LeakyReLU; output layer: linear
The quantity of neurons in every layer	(512, 256, 128, 1)

Table 4. The WGAN-GP model's hyperparameters that were set for our studies.

Parameter	Values
Enhancer	RMSprop (lr = 0.00001)

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207051 |

Quantity in batch	128
Number of epochs	50,000
Loss function	Wasserstein loss
parameter for GP regularization	10

V. RESULTS

We used precision and loss measures within the aerial video library on YouTube to assess our method. To assess the potency of classification for every action category, we also examined the confusion matrix. Our three experiments were as follows:

1. Baseline: CNN-LSTM model (accuracy: 71.21%)
2. Traditional data enhancement + CNN-LSTM model (accuracy: 94.58%)
3. Proposed model: WGAN-based feature augmentation combined with fundamental video editing + CNN-LSTM model (accuracy: 97.83%)

Our suggested framework outperformed state-of-the-art methods, including MFN-3D + GAN (68.2% accuracy) and FCN (86.63% accuracy). Table 5 summarizes the results.

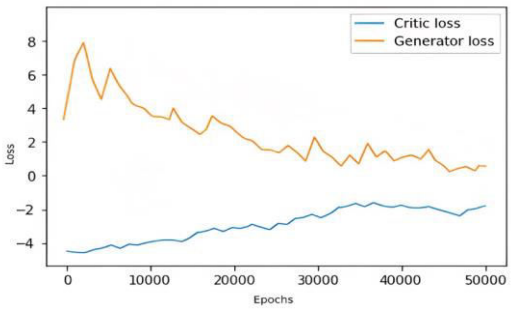

Figure 5. The WGAN-GP model's discriminator and generator faults.

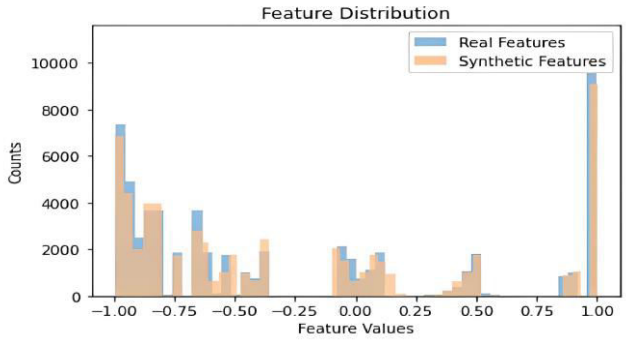

Figure 6: A bar graph illustrating the initial and artificial features.

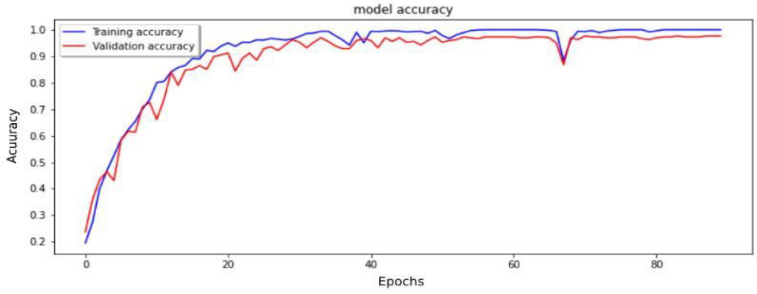

Figure 7: shows the accuracy at the two validation and training phases.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207051 |

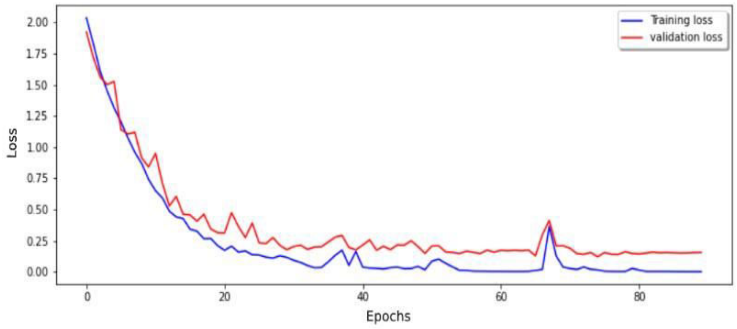

Figure 8: The error that occurred throughout the validation and training phases.

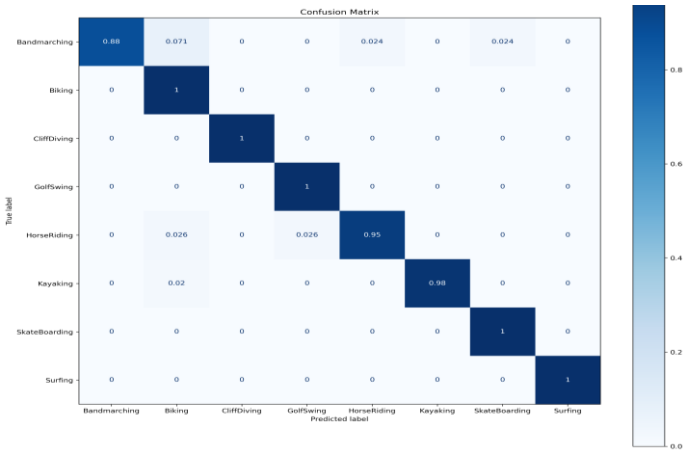

Figure 9 A table that shows the quantity of correct and incorrect predictions for each.

VI. CONCLUSION

To ensure that overcome certain data problems, we created a Hybrid GAN-Based CNN-LSTM technique to identify operations in the air footage. Our principal contributions are:
1. A new way to increase data.
2. Using GANs to create features.
3. Improved accuracy and efficiency.
However, our approach has some drawbacks, such as:
1. High memory usage.
2. Long training times.
3. Potential video quality issues.
In the future, we plan to:
1. Focus on segmenting space and time in videos.
2. Make the model more resistant to attacks.
3. Optimize processing speed, accuracy, network complexity, and computer resources.

REFERENCES

1. Gohari, A.; Ahmad, A.; Rahim, R.; Supa'at, A.; Abd Razak, S.; Gismalla, M. Involvement of Surveillance Drones in Smart Cities: A Systematic Review. IEEE Access 2022, 10, 56611–56628.
2. Mohd Daud, S.; Mohd Yusof, M.; Heo, C.; Khoo, L.; Chainchel, S.M.; Mahmood, M.; Nawawi, H. Applications of drone in disaster management: A scoping review. Sci. Justice 2022, 62, 30–42
3. Penmetsa, S.; Minhuj, F.; Singh, A.; Omkar, S.N. Autonomous UAV for suspicious action detection using pictorial human pose estimation and classification. Elcvia Electron. Lett. Comput. Vis. Image Anal. 2014, 13, 18–32.
4. Sultani, W.; Shah, M. Human action recognition in drone videos using a few aerial training examples. Comput. Vis. Image Underst. 2021, 206, 103186.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207051 |

5. Mumuni, A.; Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. Array 2022, 16, 100258.
6. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014.
7. Yacoob, Y.; Black, M.J. Parameterized modeling and recognition of activities. In Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 4–7 January 1998; pp. 120–127.
8. Ke, Y.; Hebert, M. Volumetric features for video event detection. Int. J. Comput. Vis. 2010, 88, 339–362.
9. Bobick, A.F.; Davis, J.W. The recognition of hu- man movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 2001, 23, 257–267.
10. Zhang, Z.; Hu, Y.; Chan, S.; Chia, L.-T. Motion context: A new representation for human action recognition. Motion context: A new representation for human action recognition.
In Proceedings of the Computer Vision—ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, 12–18 October 2008; Part IV, pp. 817–829
11. Efros, A.A.; Malik, J. Recognizing action at a distance. In Proceedings of the Ninth IEEE International Conference on Computer Vision—ICCV'03, Nice, France, 13–16 October 2003; Volume 2, p. 726.
12. Willems, G.; Tuytelaars, T.; Van Gool, L. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In Proceedings of the Computer Vision—ECCV, Marseille, France, 12–18 October 2008; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5303, pp. 650–663.
13. Scovanner, P.; Ali, S.; Shah, M. A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the 15th ACM International Conference on Multimedia, Augsburg, Germany, 24–29 September 2007; pp. 357–360.
14. Dollar, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior recognition via sparse spatio-temporal features. In Proceedings of the 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2005; pp. 65–72.
15. Laptev, I. On Space-Time Interest Points. Int. Comput. Vis. 2005, 64, 107–12.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details