iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37355

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Semi-Supervised Image-to-Video Adaptation for Video Action Recognition

Rohan Munshi[1], Kalpataru Podder[2], Akshay Jain[3], Abhishek Dey[4], S. Aarthi[5]
[1,2,3,4]*B-Tech, Computer Science, SRM Institute of Science and Technology, Tamilnadu, India*
[5]*Assistant Professor, Computer Science, SRM Institute of Science and Technology, Tamilnadu, India*

Abstract: ***Given a sequence of images video, the errand of activity acknowledgment is to distinguish the most same activity among the activity arrangements learned by the framework. Such human activity acknowledgment depends on proof assembled from recordings. It has a parcel of use including reconnaissance, video ordering, biometrics, telehealth, and human-PC interaction. Vision-based human action acknowledgment is tormented by various provokes because of reading changes, impediments, variety in execution rate, camera movement, and foundation clutter. In this study, we give an outline and report of the current techniques dependent on their capacity to deal with these difficulties just like how these strategies can be summed up and their capacity to distinguish irregular actions. Such precise arrangement can encourage specialists to detect the adequate ways available to manage everything about moves visaged and their limitations. In expansion to this, we additionally recognize the open datasets and the difficulties presented by them. From this review, we tend to reach determinations identifying with anyway well a test has been settled, and that we decide potential examination territories that need more work.***
Keywords: Action video recognition, convolutional neural network (CNN), two-stream, video super-resolution.

I. INTRODUCTION

Realizing human activities from a video stream is a testing task and has gotten significant consideration from the PC vision inquire about network as of late. Breaking down a human [1] activity isn't just a matter of displaying examples of movement of various pieces of the body, rather, it is additionally a depiction of an individual's expectation, feeling and considerations. Thus, it has turned into a vital part in human conduct examination and comprehension, which are basic in different areas including reconnaissance, apply autonomy, social insurance, video seeking, human-PC communication, and so on. It is not same as still picture classification, video information contains worldly data which assumes a critical job in real life acknowledgment. Moreover, video information incorporates normal information enlargement, for example jittering[2] for video outline classification. One instinctive answer for the scale fluctuation issue is performing SR of the low-goals recordings preceding activity acknowledgment. Action recognition task involves the checking of various actions from video clips where the action may or may not be performed throughout the entire duration of the video. This seems like a natural extension of image classification tasks to various frames and then aggregating the predictions from each frame. Though the stratospheric success of deep learning architectures in image classification (ImageNet), progress in architectures for video classification and representation learning has been slower.

Table I
The Psnr And Action Recognition Accuracy Using Different Sr Methods On Ucf101 Dataset (Split1)

Method	PSNR (dB)		Recognition Accuracy (%)	
	2×	4×	2×	4×
Bi cubic	36.54	29.54	92.92	88.42
VDSR [6]	40.76	32.00	93.32	86.45
Video SR [7]	32.16	29.97	93.52	86.46
OR video	-	-	93.49	

PSNR shown here is the average PSNR computed on some randomly sampled video frames. It is worth noting that the PSNR of Video SR method is not accurate due to the sub-pixel shift between HR frames and SR frames. However, this shift does not seem to influence the recognition performance. Profound learning-based human activity acknowledgment [3] approaches connected to various datasets are talked about in Section 3. Areas 4 and 5 recommend potential research openings and give an end, separately.

II. RELATED WORK

Automatic inspection, e.g., in manufacturing applications. Assisting humans in identification tasks, e.g., a process of species identification system

Controlling processes, e.g., an industrial robot. Detection of events, e.g., for visual surveillance or people counting. Changes with each other, e.g., as the input to a device for computer-human interaction. Modeling objects or nature,[4] e.g., medical image analysis or topographical modelling. In appearance based methods, less accurate of features description because of whole image consideration. In geometric based methods, the features like distance between eyes, face length and width, etc., are considered which not provides optimal results.

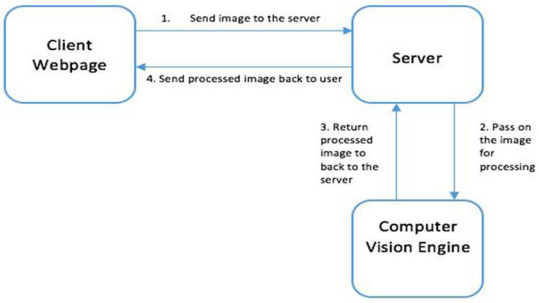


Figure 1: The relation between client webpage and server system.

Recognizing Human Actions: A Local SVM Approach (2004)

1) Use local space-time features to show video sequences that contain actions.
2) Classification is done via an SVM. Results are also computed for KNN [5] for comparison.
3) Christian Schuldt, Ivan Laptev and Barbara Caputo (2004).

Unsupervised Learning of Human Action Categories with the use of Spatial-Temporal Words (2008)

a) Unsupervised way to deal classifying actions that occur in video.
b) Uses pLSA (Probabilistic Latent Semantic Analysis) to learn a model.
c) uan Carlos Niebles, Hongcheng Wang and Li Fei-Fei (2008).

Brightness gradients calculated for each interest point 'cube'. Image gradients computed for each cube (at different scales of smoothing). Gradients concatenated to form feature vector. Length of vector: Vector is projected to lower dimensions using PCA.

III. DATASETS

With the improvement of human activity acknowledgment innovation, a wide range of kinds of datasets have been arranged and discharged as of late. These datasets are generally utilized for test purposes to assess the execution and precision of existing/new methodologies and to guarantee fitting examination with different methodologies. For the most part, profound [6] learning can be connected to various sorts of datasets with crude information. What's more, the multifaceted nature of the systems might be dictated by the distinctive sorts of datasets. For instance, solitary perspective information may require less advance than various perspective information, which needs to produce numerous systems to acquire the final yield.

1) *Single Viewpoint Datasets:* Video dataset with a few thousand instances.25 people each: perform 6 different actions. Walking, jogging, running, boxing, hand waving, clapping in 4 different scenarios Outdoors, outdoors w/scale variation, outdoors w/different clothes, indoors (several times). Backgrounds are mostly free of clutter. Only one person performing a single action per video.

Figure 2: Human action recognition by a typical single viewpoint.

2) *Space-time Interest Point:* Training set- Set of videos in which a single person is performing a single action. Videos are unlabeled. Test set (relaxed requirement): Set of videos which can contain multiple people performing multiple actions simultaneously.

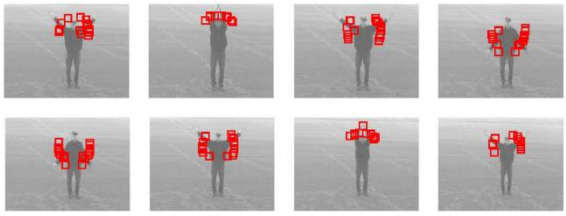


Figure 3: Space-time interest points where each movement is being captured.

Brightness gradients calculated for each interest point 'cube'. Image gradients [7] computed for each cube (at different scales of smoothing). Gradients concatenated to form feature vector.
Length of vector: [# of pixels in interest "point"] * [# of smoothing scales] * [# of gradient directions].
Vector is projected to lower dimensions using PCA.

IV. ANALYSIS

Action parsing in videos with complicated scenes is a noteworthy however difficult task in pc vision.
In this paper, we tend to propose a generic 3D convolutional neural network in a very multi-task learning manner for effective Deep Action Parsing (DAP3D-Net) in videos.
Particularly, within the coaching section, action localization, classification and attributes learning will be conjointly optimized on our appearance-motion information via DAP3D-Net.
For associate degree forthcoming take a look at video, we will describe every individual action within the video at the same time as: wherever the action happens, What the action is and the way the action is performed.
To well demonstrate the effectiveness of the planned DAP3D-Net,[8] we tend to conjointly contribute a brand new Numerous-category Aligned artificial Action dataset, i.e., NASA, that consists of two hundred, 000 action clips of
more than 300 categories and with 33 pre-defined action attributes in two hierarchical levels (i.e., low-level attributes of basic body part movements and high-level attributes related to action motion). Approaches focuses on 2D images.
The proposed system [9] describes each individual action as-Where the action occurs, What the action is, How the action is performed, In other words, it shows accurately localize, categorize and describe multiple actions in realistic videos., Automatically parse the action in videos.

A. *Advantages of Proposed System*

1) 3D Images are supported
2) Object detection, Scene Attributes Learning or Action Recognition are done as unified model.

The Working model consists of 5 modules as follows-

a) *Module 1*-video Upload
From the client we upload the video file to server. The video file contains the information about the action to be detected. Once uploaded the server stores the file the server hard disk.
b) *Module 2*-Frame Separation
In this module, we retrieve one by one from the uploaded videos. Each retrieved frame will be stored in the server for further analyzing.
c) *Module 3*-Background and Noise Removal
In this module, we retrieve one by one from the uploaded videos. Any noise (not sound) (any unwanted dot or line) will be removed from the frame images.
d) *Module 4*-Detect Action
Using Deep Learning Algorithm,[10] detect the action which is available in the uploaded videos.

V. RESULT

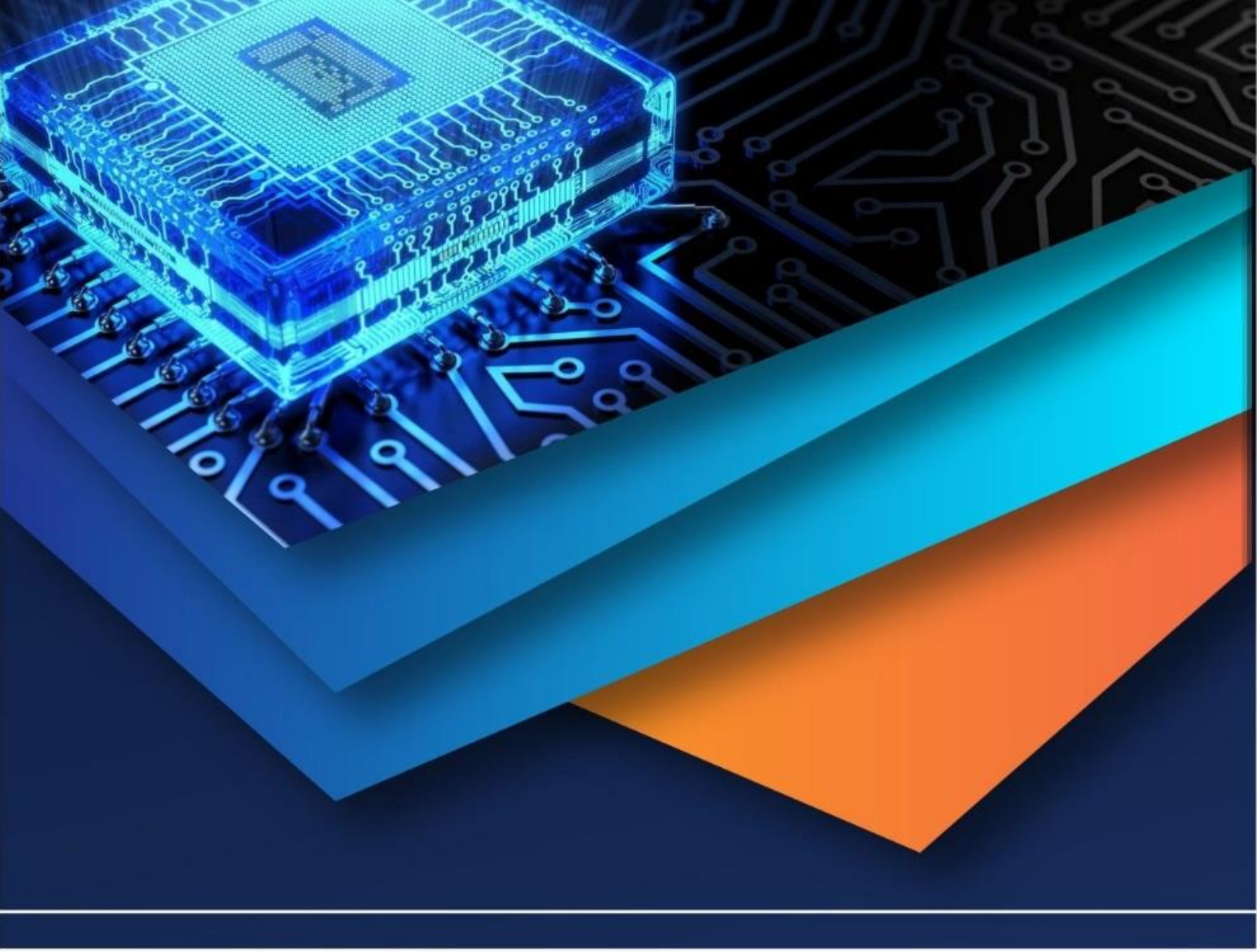
Video action recognition accuracy are summarized in Table III including the bicubic baseline, VDSR, Video SR method and our method. It is observed that using our training strategy, the recognition accuracy of the optical flow stream exceeds other SR methods as well as of the bicubic baseline. temporal profiles of the 'Body-Weight-Squats g02 c04' video clip, which demonstrates a better temporal consistency of our proposed SR network. We perform late fusion on spatial stream (using bicubic as it is better than the other SR methods) and temporal stream (using our proposed method) by the weight of 1:1.5 which is suggested by [8]. We achieve 88.95% accuracy as the final result which outperforms the other SR methods as well as the bicubic baseline.

VI.CONCLUSION

Deep learning techniques have recently been introduced in the video-based human action recognition research area. They have been widely used in other areas, such as speech recognition, language processing and recommendation systems etc. There are many advantages to hierarchical statistical approaches, such as raw data input, self-learnt features and a high-level or complex action recognition, hence, deep learning techniques have received much interest. Based on these advantages, researchers could design a real-time, adaptive and high performing recognition system. However, these approaches also have several drawbacks, such as the need to generate large datasets, the performance depends on the scale of the network weights and hyper-parameter tuning is non-trivial etc. In this review, we presented techniques mainly focusing on developments in deep learning over the past five years. Many investigations have been conducted to deal with different types of datasets. For single/multiple viewpoint approaches, the inputs are normally frames, so researchers have performed 3D convolution operations to add the temporal information in order to recognize videos. Aditionally some of the approaches.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.
[2] J. Hager, P. Ekman, and W. Friesen, "Facial action coding system. salt lake city, ut: A human face," ISBN 0-931835-01-1, Tech. Rep., 2002.
[3] Z. Zhang, "Feature-based facial expression recognition: Sensitivit analysis andexperimentswithamultilayerperceptron,"International journal of pattern recognition and Artificial Intelligence, vol.13, no. 06, pp. 893–911, 1999.
[4] G. Guo and C. R. Dyer, "Simultaneous feature selection and classifier training via linear programming: A case study for face expression recognition," in Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 1. IEEE, 2003, pp. I–346.
[5] M. F. Valstar, I. Patras, and M. Pantic, "Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops. IEEE, 2005, pp. 76–76.
[6] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06). IEEE, 2006, pp. 149– 149.
[7] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.
[8] C. Padgett and G. W. Cottrell, "Representing face images for emotion classification," Advances in neural information processing systems, pp. 894–900, 1997.
[9] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," IEEE Transactions on pattern analysis and machine intelligence, vol. 21, no. 10, pp. 974–989, 1999.
[10] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu, "A principal component analysis of facial expressions," Vision research, vol. 41, no. 9, pp. 1179–1208, 2001.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY