The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6655091

LEARNING SPATIO-TEMPORAL FEATURE EXTRACTION USING RESIDUAL FRAMES WITH NEURALNETWORKS FOR HUMAN ACTION RECOGNITION

C.INDHUMATHI[1], Dr.V.MURUGAN[2] and Dr.G.MUTHULAKSHMI[3]

[1]Department of Computer Science and Engineering Manonmaniam Sundaranar University Abishekapatti,Tirunelveli,TamilNadu,India.Email:indhuinfo2013@yahoo.com
[2]Department of Computer Science Government Arts and Science College Kadayanallur,Tenkasi, Tirunelveli,TamilNadu,India.Email:smv.murugan@gmail.com
[3]Department of Computer Science and Engineering Manonmaniam Sundaranar University Abishekapatti,Tirunelveli,TamilNadu,India.Email:lakshmi_me05@yahoo.co.in

Abstract

In recent times the growth of machine learning and artificial intelligence algorithms help to expand the use of image and video processing. The usage of different algorithms is applicable in various fields such as content-based video recognition, video surveillance, assistive living, autism care, and gaming. HAR (Human Action Recognition) highly demands efficient computation. This research proposed a method for selecting residual frames and keyframes toeliminate redundant information from videos. This method combines the extraction of spatial and temporal features. These features were extracted using the VGG16 (Visual Geometry Group) network and classified using Multi SVM classifier. The proposed research method was tested on HMDB51 and UCF101 datasets. The result of the proposed method achieved an accuracy of 85.6% and 98.71% onHMDB51andUCF101datasetsrespectively.
Keywords: Spatial features, Temporal features, Keyframes, Residual frames, VGG16

I. INTRODUCTION

Human actions are recognized from still images and videos. It is the main part in content-based retrieval system. The ranges of actions lie from slow to fast. The action features from video as spatio-temporal features. Hand-crafted features and deep features are extracted from the last several decades. Several pre-trained models are also used to extract features and classify actions. There are various datasets for action recognition which is classified as daily activity, sports activity and so on. One such dataset is introduced in [1] for still image Action Recognition (STAR), which contains over 1M images across 50 different human body-motion action categories. This is the largest dataset for action recognition in still images.

To improve the belief network, multi-scale input data, spatiotemporal Deep Belief Network (DBN), and different pooling strategiesare analysed [2]. Moreover, a human sports action recognition model has also been developed based on particular spatiotemporal features. In [3], supervised action recognition model has been introduced for dark region. This work has used Action Recognition in the Dark (ARID) dataset for testing which is meant for recognizing actions in the dark.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6655091

K. Hara et al. [4] developed a Convolutional Neural Network (CNN) based approach to recognize human events in video and image data.Although the overall system is robust, the issue is the weak model of the system. L. Zhang [5] has presented a two-level neural network-based approach for human action recognition. For the first level, CNNs are trained to provide information using video to an event, which understands the important content of the video. At the second level, they use a Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) based technique to handle both temporal and spatial information.

In [6], Q. Meng et al. have developed the SVM classification based feature extraction method for action classification. But this work suffers from low accuracy. Shapovalova et al. [7] has developed a combination of spatiotemporal features with Bag of Words (BoW) to model sub-regions of the video. Wang et al. [8] employed the standard Bag of Features (BoF) to evaluate four dense trajectory features (HOG, HOF, MBH, and motion trajectory) on four different datasets KTH, UCF sports, YouTube, and Hollywood.

Aly and Sayed [9] has introduced a combination of global and local Zernike moment features extracted from temporal frames.Khan et al. [10] have presented a combined Deep Neural Network (DNN) and handcrafted features based method for human action recognition. The system complexity and computational time are high.

In this method, the video is divided into scenes. In each scene, the spatial features are extracted from keyframes and the temporal features are extracted from intermediate frames. Residual frames are created from the intermediate frames using Frame Differencing (FD) method. VGG16 is used for extracting both spatial and temporal features. Both the features are combined for each scene and classified using multiclass SVM classifier.

The main contributions of this paper include:

- The video is divided into scenes of different sizes.
- From each scene, spatial features are extracted from the first frame. Temporal features are extracted from the intermediate frames.
- VGG16 is used for spatial feature extraction. Multiple VGG16 networks are used for temporal feature extraction.
- Both the features are fused and classified using multiSVM classifier.

The remaining of the chapter is organized as follows: Section 2 discusses some related works that are used for comparing the proposed method. Section 3 elaborates the architecture and methodologies used in proposed method. Section 4 demonstrates experimental results and analysis. Section 5 gives conclusion and future scope.

II. Related Works

This section discusses some most recent methods in human action recognition. In [11], contrastive learning are generalized to a wider set of transformations, and their compositions, for which either invariance or distinctiveness is sought. This method is named as Generalized Data Transformation (GDT).

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6655091

Humam Alwasselet al. has implemented action recognition for audio enabled video data. Aself-supervised Cross-Modal Deep Clustering (XDC) method has been developed that leverages unsupervised clustering in audio as a supervisory signal for video. This method has used the semantic correlation and the differences between audio and video.

To improve spatiotemporal 3D CNNs, some analysis are done in [13].It is examined that whether large-scale video datasets will help to improve spatiotemporal 3D CNNs in terms of video classification accuracy. From the analysis, it is found that a carefully annotated dataset (e.g., Kinetics-700) effectively pre-trains a video representation for a video classification task. In [14], actions are recognized in dark videos. This method creates a new dataset for its testing to bridge the gap of the lack of data.

In [15], 3D rendering tools has been used to generate a synthetic dataset of videos, and show that a classifier trained on these videos can generalize to real videos. 3D convolution is combined with late temporal modeling for action recognition [16]. For this, the conventional Temporal Global Average Pooling (TGAP) layer is replaced at the end of 3D convolutional architecture with the Bidirectional Encoder Representations from Transformers (BERT) layer in order to better utilize the temporal information with BERT's attention mechanism.

A multi-viewpoint outdoor action recognition dataset collected from YouTube has been presented in [17]. Another dataset is also created for their research. This dataset is useful to many research areas including action recognition, surveillance etc. A Meta-Contrastive Network (MCN) [18] has been developed, which combines the contrastive learning and Meta learning, to enhance the learning ability of existing self-supervised approaches. This method contains two training stages based on Model-Agnostic Meta Learning (MAML), each of which consists of a contrastive branch and a meta-branch.

A feature selection method named Poisson distribution along with Univariate Measures (PDaUM) has been developed in [19]. In this method, few of fused CNN features are irrelevant, and few of them are redundant that makes the incorrect prediction among complex human actions. Therefore, this method selects only the strongest features which are given to the Extreme Learning Machine (ELM) classifier.

In our previous work [20], Adaptive motion Attentive Correlated Temporal Feature (ACTF) is incorporated for temporal feature extractor. The temporal average pooling in inter-frame is used for extracting the inter-frame regional correlation feature and mean feature.

Another work recognizes actions using key frame segmentation method [21]. In that method, features from key frames are extracted using MultiFiber Network (MFNet) and the whole video is used for temporal feature extraction using Long Short Term Memory (LSTM) network.

In this paper, only residual frames are used for temporal feature extraction. Since, key frames are used for spatial feature extraction; it is redundant if we use the same for temporal feature extraction. Only the motion information is necessary for temporal features.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6655091

III. Proposed Methodology

Human action recognition in videos consists of spatial feature extraction, temporal feature extraction and classification. The proposed method gives significant focus on temporal feature extraction.The proposed system architecture is shown in Fig. 1. Initially, the video in the dataset is divided into Group of Pictures (GOP) which can also be named as scenes. Each GOP consists of keyframes and intermediate frames. The keyframes are the representative frames for each GOP. The intermediate frames are the frames between two consecutive keyframes. The keyframes are given to spatial feature extraction. From the intermediate frames, residual frames are created using Frame Differencing (FD) method. Temporal features are extracted from residual frames. Both features are concatenated using multiclass SVM classifier.

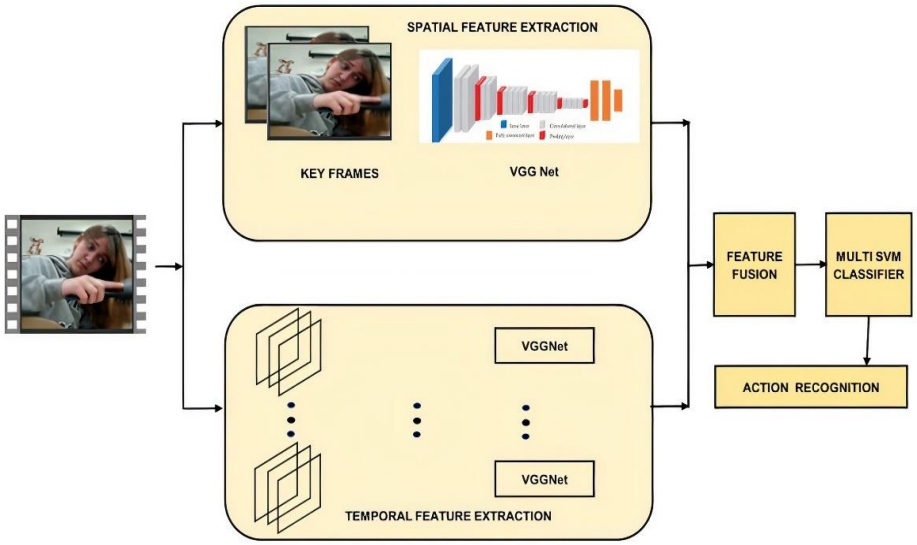

Fig. 1 Proposed System Architecture

The classification task is divided into two main stages. The first stage is to extract spatial features based on scene content. The second stage is to build motion representation. In the first stage, only keyframes are selected from the video. And the fine-tuned VGG-Net is transferred as the basic architecture to extract spatial features from full-connected layers. In the second stage, the main steps include: (1) Frame Differencing (2) Pooling. Finally, spatial features and temporal features are stacked for feature fusion, and a multiclass SVM is used as a classifier for action recognition.

Spatial Feature Extraction

In Spatial feature extraction, only keyframes are extracted using Scene Change based Segmentation (SCS) [22]. According to this algorithm, the input video is divided into GOP. From each GOP, a single frame is selected as keyframe. In every video, the first and last frames are selected as keyframes by default. From the intermediate frames, keyframes are selected using SCS Algorithm.

ISSN 1533-9211

DOI 10.5281/zenodo.6655091

Spatial features are extracted from these keyframes using a simple VGG16 network. The VGG-Net used in this work has 19 weight layers, which comprises five blocks of convolutional layers and three fully connected layers, where each block is followed by one pooling layer [23]. In this architecture, the input size of images is resized to 224 × 224 × 3, and the size of the last fully connected layer is modified with the number of classes to be classified. The architecture of VGG-Net used in this paper is shown in Fig. 2.

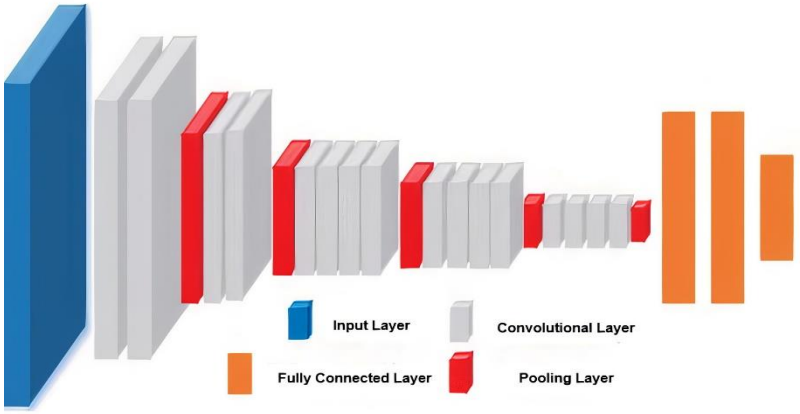

Fig. 2 VGG16 Network Architecture

Temporal Feature Extraction

In temporal feature extraction, the motion information between keyframes and the intermediate frames are obtained by FD method. It is calculated as the absolute difference between two frames f_1 and f_2:

$$FD = |f_1 - f_2| \quad (1)$$

As discussed in the previous subsection, each GOP has two keyframes (the first and last frame in each GOP). The difference between the first keyframe and the intermediate frame makes the residual frame. The group of residual frames isgiven to VGG16 network to create temporal features. The size of temporal feature for a scene depends on the number of residual frames in each scene. Similarly, the size of spatial feature depends on the number of keyframes in each scene (obviously, it is 1). Hence, the average of temporal features is obtained for every scene in order to make compatible with spatial features.

Spatio-temporal Feature Extraction

The method of extracting spatial and temporal features is integrated to a single method which is shown in Algorithm 1. Initially, the dataset is split into training and testing set. In each set, the following algorithm is applied to extract Spatio-temporal features. The video in the dataset is divided into frames. The first frame is set as keyframe. The next keyframe is identified from the next sequence of frames. For this, PCC is calculated between keyframe and the next frames. If the value of PCC is greater than a threshold, we identified the next keyframe. Once it is identified, the residual frames are calculated by finding the difference between the two keyframes.

ISSN 1533-9211

DOI 10.5281/zenodo.6655091

Next, the residual frames are given to pre-trained VGG16 network. The same VGG16 architecture (Fig. 2) used in spatial feature extraction is used here. The features from the last FC layer are taken and the average of it is calculated. The obtained feature is the temporal feature extracted for one scene. Figure 3 illustrates the spatial temporal feature extraction of a scene.

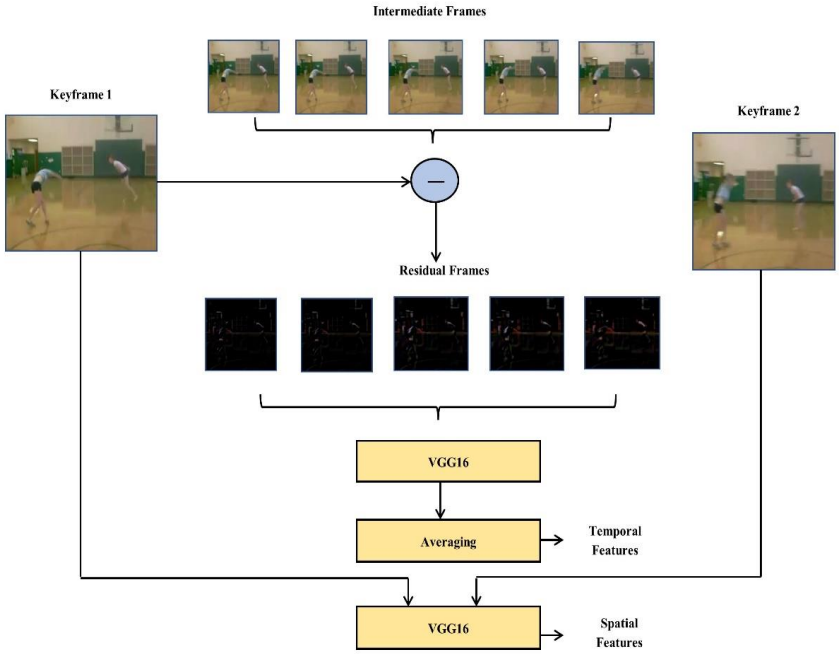

Fig. 3 Illustration of Spatial and Temporal Feature Extraction for a scene

Once all the keyframes are identified, the final residual frames are calculated using the last frame in the sequence. Finally, the spatial features are extracted by giving all the keyframes as the input to VGG16 network.

Let the spatial features extracted from keyframes are given as $f_{spatial}$ and the temporal features are given as $f_{temporal}$. Since the number of frames given to each video is same as for spatial and temporal, the output features extracted from both VGGNet are of similar size. Hence it is easy to fuse both the features column-wise. The spatio-temporal feature extraction method is shown in Fig. 3

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6655091

Algorithm: 1 Spatio-Temporal Feature Extraction

Input: Dataset, VGG Network Model
Output: Spatio-temporal features
Steps: 1. For each sequence I in Dataset
1.1 Set Keyframe K_f={first_frame}
1.2 f_k=first_frame
1.3 For each frame f in I
1.3.1 Calculate PCC between f_k and f using

$$PCC = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(f_k(i,j) - f_k^m)((f(i,j) - f^m)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(f_k(i,j) - f_k^m)^2(f(i,j) - f^m)^2}}$$

1.3.2 If PCC>keyframe_threshold
1.3.2.1 Concatenate f into K_f
1.3.2.2 For each frame between f_k and f
Calculate frame difference between f_k and f using

$$FD = |f_k - f|$$

Concatenate FD to FD_f
1.3.2.3 End
1.3.2.4 Give FD_f to pre-trained VGG16 Network Model
1.3.2.5 Get the features from the last FC layer
1.3.2.6 Find the average of the obtained features and concatenate to $f_{temporal}$
1.3.2.7 f_k=f
1.3.3 Else
Go to next frame
1.4 End
1.5Concatenate last_frame to K_f.
1.6 For each frame between f_k and last_frame
1.6.1 Calculate frame difference between $f_k$and last_frame using

$$FD = |f_k - last\_frame|$$

1.6.2Concatenate FD to FD_f
1.7 End
1.8 GiveK_f to pre-trained VGG16 Network Model
1.9 Get the features from the last FC layer $f_{spatial}$
1.10 Concatenate $f_{spatial}$ and $f_{temporal}$ to create $f_{spatio\_temporal}$
2. End

In the above algorithm, the input to VGG16 for spatial feature extraction is the keyframesk_f. Similarly, the input to VGG16 for temporal feature extraction is residual framesFD_f. The keyframe_threshold is set to 0.8 by analyzing various values of PCC which are discussed in Section 4. If the number of keyframes for a sequence is 10, then the size of k_f is 10.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6655091

Obviously, the size of FD_f will be 9. Hence, the last frame differencing frame is again concatenated to match the size of spatial features.

Feature Fusion and Classification

The spatial and temporal features thus obtained are pooled to get a single feature for each sequence. The spatial and temporal features are concatenated column-wise as

$$f_{spatio-temporal} = [f_{spatial} f_{temporal}] \quad (2)$$

The size of $f_{spatio-temporal}$ is r x c where r is the number of sequences in the dataset and c is the 8192. Finally for each video sequences, the obtained spatial and temporal representations are then concatenated to train a non-linear support vector machine classifier with a multichannel $\chi 2$ kernel. A support vector machine is a discriminative classifier formally defined by a separating hyperplane.

IV. Experimental Results

$$Ac_r = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100 \quad (3)$$

Where Ac_r is the accuracy rate, $T_p$is the True Positive rate, T_n is the True Negative rate, $F_p$is the False Positive rate and F_n is the False Negative rates.

The proposed action recognition method is executed in Nvidia Titan X GPU. Table 1 show the hyper-parameters used to train and test VGG16 network.

Table 1 Hyperparameters of VGG16 Network Model

Hyperparameter	Value
Mini Batch Size	16
Dropout	0.8
Learning Rate	0.001
Optimizer	Adam

The efficacy of the proposed method can be proved only when it is compared with other methods. Table 2 shows the comparison of proposed method with recent methods. The proposed action recognition method is compared with recent methods [11-19] that are discussed in Section 2.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6655091

Table 2 Accuracy Comparison of Proposed Method with Recent Methods on HMDB51 and UCF101 Datasets

Dataset/Method	HMDB51	UCF101
GDT [11]	72.8	95.2
XDC [12]	65.1	94.2
Kataoka et al. [13]	69.4	92.9
Yuecong Xu et al. [14]	63.8	-
Matthews et al. [15]	83	-
Kalfaoglu et al. [16]	85.1	98.69
Perera et al. [17]	72.7	-
Lin et al. [18]	54.8	85.4
Khan et al. [19]	81.4	-
Proposed Methodology	85.6	98.71

From Table 2, it is observed that the proposed method achieves 85.6% and 98.71% accuracy on HMDB51 and UCF101 datasets respectively. In HMDB51 dataset, the proposed method achieves 0.5% higher accuracy than Kalfaoglu et al's method [15]. Similarly, in UCF101 dataset, it achieves 98.71% higher accuracy than Kalfaoglu et al.'s method [15].

Ablation Study

The proposed method is analyzed for various PCC threshold values. The value of PCC lies between 0 (no correlation) and 1 (highly correlated). For splitting the video into scenes, PCC is used in this work. The value of PCC is varied from 0.3 to 0.8. Remaining values are left out, as it is not useful for our work. A sample video is taken for this analysis whose scenes are identified manually. The number of frames for each scenes obtained by the proposed method is compared with manual scene identification for various values of PCC which is given in Table 3.

Table 3 Comparison of Manual and PCC based Scene Identification

Scenes	Manual (No. of frames)	PCC Based (No. of frames)					
		Threshold = 0.3	Threshold = 0.4	Threshold = 0.5	Threshold = 0.6	Threshold = 0.7	Threshold = 0.8
Scene 1	10	1	2	4	6	8	10
Scene 2	12	1	2	5	7	10	12
Scene 3	9	1	2	3	5	7	9
Scene 4	15	1	3	5	8	12	14

From the Table 3, it is clear that the number of frames in every scene is correctly identified when keyframe threshold is set to 0.8.

The proposed method is also tested with AlexNet instead of VGG16. The other networks suffer from heavy computation time. Hence, only AlexNet is used for analysis. Table 4 shows

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6655091

the accuracy achieved by the proposed method for both the networks with the same hyperparameters listed in Table 1.

Table 4 Accuracy Obtained by the Proposed Method for different Network Models

Network Model	HMDB 51	UCF101
AlexNet	84.7	96.8
VGG16	85.6	98.71

From Table 4, it is evident that the proposed method achieves high accuracy for VGG16 network.

V. Conclusion

In the proposed method, the videos in the human action recognition dataset are divided into scenes in which the scene consists of different number of frames. In each scene, the keyframes are used extract spatial features and intermediate frames are used to extract temporal features. The intermediate frames are converted to residual frames using keyframes. Both the spatial and temporal features are concatenated and classified using multiSVM classifier. This method is tested on two publicly available datasets UCF101 and HMDB51. It achieves 85.6% and 98.71% accuracy on HMDB51 and UCF101 datasets respectively which are higher than recent methods.

References:

[1] Safaei, M., Balouchian, P. and Foroosh, H., 2020, April. UCF-STAR: A Large Scale Still Image Dataset for Understanding Human Actions. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 03, pp. 2677-2684).

[2] Guo, Y. and Wang, X., 2021. Applying TS-DBN model into sports behavior recognition with deep learning approach. The Journal of Supercomputing, pp.1-17.

[3] Hira, S., Das, R., Modi, A. and Pakhomov, D., 2021. Delta Sampling R-BERT for Limited Data and Low-Light Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 853-862).

[4] Hara, K.; Kataoka, H.; Satoh, Y. Can spatiotemporal 3d CNN's retrace the history of 2d CNN's and imagenet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6546–6555

[5] Banu, J. F., Muneeshwari, P., Raja, K., Suresh, S., Latchoumi, T. P., & Deepan, S. (2022, January). Ontology Based Image Retrieval by Utilizing Model Annotations and Content. In 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 300-305). IEEE.

[6] Meng, Q.; Zhu, H.; Zhang, W.; Piao, X.; Zhang, A. Action Recognition Using Form and Motion Modalities. ACM Trans. MCCA 2020

[7] Garikapati, P. R., Balamurugan, K., Latchoumi, T. P., & Shankar, G. (2022). A Quantitative Study of Small Dataset Machining by Agglomerative Hierarchical Cluster and K-Medoid. In Emergent Converging Technologies and Biomedical Systems (pp. 717-727). Springer, Singapore.

[8] Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. 103(1), 60–79 (2013)

[9] Latchoumi, T. P., & Parthiban, L. (2022). Quasi oppositional dragonfly algorithm for load balancing in cloud computing environment. Wireless Personal Communications, 122(3), 2639-2656.

The Seybold Report
ISSN 1533-9211
DOI 10.5281/zenodo.6655091

[10] Sharif M, Akram T, Raza M, Saba T, Rehman A (2020) Hand-crafted and deep convolutional neural network features fusion and selection strategy: an application to intelligent human action recognition. Appl Soft Comput 87:105986
[11] Karnan, B., Kuppusamy, A., Latchoumi, T. P., Banerjee, A., Sinha, A., Biswas, A., & Subramanian, A. K. (2022). Multi-response Optimization of Turning Parameters for Cryogenically Treated and Tempered WC–Co Inserts. Journal of The Institution of Engineers (India): Series D, 1-12.
[12] Humam Alwassel, Bruno Korbar, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In NeurIPS, 2020
[13] Pavan, V. M., Balamurugan, K., & Latchoumi, T. P. (2021). PLA-Cu reinforced composite filament: Preparation and flexural property printed at different machining conditions. Advanced composite materials.
[14] Yuecong Xu , Jianfei Yang , Haozhi Cao , Kezhi Mao , Jianxiong Yin and Simon See, "ARID: A Comprehensive Study on Recognizing Actions in the Dark and A New Benchmark Dataset", arXiv preprint arXiv: 2006.03876
[15] Latchoumi, T. P., Kalusuraman, G., Banu, J. F., Yookesh, T. L., Ezhilarasi, T. P., & Balamurugan, K. (2021, November). Enhancement in manufacturing systems using Grey-Fuzzy and LK-SVM approach. In 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT) (pp. 72-78). IEEE.
[16] . Kalfaoglu, M.E., Kalkan, S. and Alatan, A.A., 2020, August. Late temporal modeling in 3d cnn architectures with bert for action recognition. In European Conference on Computer Vision (pp. 731-747). Springer, Cham.
[17] Utilizing scratch to create computational thinking at school with artificial intelligence Kumari, M.K., Latchoumi, T.P., Kalusuraman, G., Chithambarathanu, M., Parthiban, L.A Closer Look at Big Data Analytics, 2021, pp. 163–193
[18] Lin, Y., Guo, X. and Lu, Y., 2021. Self-supervised video representation learning with meta-contrastive network. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8239-8249).
[19] Khan, M.A., Zhang, Y.D., Khan, S.A., Attique, M., Rehman, A. and Seo, S., 2020. A resource conscious human action recognition framework using 26-layered deep convolutional neural network. Multimedia Tools and Applications, pp.1-23.
[20] Indhumathi, C., Murugan, V. and Muthulakshmii, G., 2021. Human Action Recognition Using Spatio-Temporal Multiplier Network and Attentive Correlated Temporal Feature. International Journal of Image and Graphics, p.2250051.
[21] Tracking system for birds migration using sensors Bhavya, B., Rajesh, T.R., Latchoumi, T.P., Harika, N., Parthiban, L.A Closer Look at Big Data Analytics, 2021, pp. 195–223
[22] Sowmyayani, S. and Rani, P.A.J., 2016. Frame differencing-based segmentation for low bit rate video codec using H. 264. International Journal of Computational Vision and Robotics, 6(1-2), pp.41-53.
[23] Simonyan K, Zisserman A. Very deep convolutional networks for large scale image recognition. In: Proceedings of International Conference on Learning Representations (ICLR), San Diego, 2015. 1–14
[24] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. Technical Report CRCV-TR-12-01, 2012.
[25] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In Proc. ICCV, 2011.