



# STHARNet: spatio-temporal human action recognition network in content based video retrieval

S. Sowmyayani<sup>1</sup> · P. Arockia Jansi Rani<sup>2</sup>

Received: 11 April 2022 / Revised: 22 September 2022 / Accepted: 6 October 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

## Abstract

Most of the needed information is easily accessible from our fingertips using the internet. The search procedure via the internet is a tough task behind the scenes. Content-Based Video Retrieval (CBVR) is one such search procedure on the internet. Human action recognition is one of them in CBVR. Even though there is research in these areas, the challenges are also partially solved. This paper also addresses the issues in human action recognition by designing an architecture named the Spatio-Temporal Human Action Recognition Network (STHARNet). The proposed STHARNet system is integrated into the CBVR system. The performance of the proposed architecture is evaluated by testing on three publicly available datasets: UCF Sports, KTH, and HMDB51. The results of the proposed architecture are encouraging and better than other recent methods.

**Keywords** Spatial features · Temporal features · Keyframes · Deep learning

## 1 Introduction

The searching procedure makes it possible for humans to easily access the necessary information. It starts from local devices and grows globally using the internet. The history of search starts with the dictionary. It slowly grows to be able to search books using alphabetic order indexing. Once electronic devices entered into human lives, the searching technology also entered into electronic devices. On local devices, search is classified as search by text and

---

✉ S. Sowmyayani  
sowmyayani@gmail.com

P. Arockia Jansi Rani  
jansimsuniv@gmail.com

<sup>1</sup> Department of Computer Science, St. Mary's College (Autonomous), Thoothukudi, Tamilnadu, India

<sup>2</sup> Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India

search by audio. On the internet, the searching procedure is further classified into two categories: search by image and search by video.

Searching by image is called Content-Based Image Retrieval (CBIR), and searching by video is named Content-Based Video Retrieval. In CBVR, the input is a short video or scene for which the user is searching. The output yields as many videos as the retrieval algorithm can retrieve that contain the given query video. There are several challenges faced by a CBVR system. A CBVR system should be designed in such a way that the proposed system should be able to adopt cropped images, compressed images, and so on. The proposed CBVR system follows supervised learning. Hence, the system is divided into training and testing phases. In both the phases, the video is divided into scenes and features are extracted. In the training phase, the extracted features are stored in a dataset. In the testing phase, the extracted feature is matched with the feature dataset.

Human action videos consist of spatial and temporal features. Spatial features are those extracted from a single frame. Temporal features are those extracted from multiple frames. It shows the motion in the video. Temporal features are the ones that make CBVR look different from CBIR. This paper develops the architecture for HAR and integrates it with the CBVR system. The main contributions of this paper include:

- Spatio-temporal human action recognition architecture is designed.
- Spatial features are extracted from keyframes.
- Temporal features are extracted from Motion Energy Image (MEI) of other frames.
- The proposed HAR architecture is integrated into CBVR system.

The remainder of the paper is organized as: Section 2 discusses some recent methods in HAR and CBVR systems. Section 3 elaborates on the proposed STHAR system with its architecture and methodology. It also shows the integration of STHAR into the CBVR system. Section 4 demonstrates the proposed STHAR-CBVR method with some experiments and analysis. Section 5 concludes the work.

## 2 Related works

This section discusses some recent methods that are used for comparing the proposed method. Convolutional Neural Network (CNN) models are used in many classification systems. CNN can be used both as a feature extractor and a classifier. Many researchers build their own CNN architecture for HAR. Some of those methods that use CNN as their base model are briefly discussed here.

A feature selection method named Poisson Distribution along with Univariate Measures (PDaUM) [7] was developed in 2021. A few of the fused CNN features are irrelevant, and a few of them are redundant, which makes the incorrect prediction among complex human actions. This method selects only the strongest features, which are then given to the Extreme Learning Machine (ELM) and Softmax for final recognition.

New attention network architecture, termed the Cascade multi-head Attention Network (CATNet) [21], has been developed for HAR. It constructs video representations with two-level attention, namely multi-head local self-attention and relation-based global attention. Starting from the segment features generated by backbone networks, CATNet first learns multiple attention weights for each segment to capture the importance of local features in a

self-attention manner. With the local attention weights, CATNet integrates local features into several global representations and then learns the second level attention for the global information in a relational fashion.

Another network has been designed that captures multimodal correlations over arbitrary timestamps [23]. This method operates as a complementary, extended network over a multimodal network. Spatial and temporal streams are fused using Shannon fusion for learning a pre-trained CNN. A long-range video may consist of spatiotemporal correlations over arbitrary times, which can be captured by forming the correlation network from simple, fully connected layers.

A new trajectory descriptor has been designed [22] based on local feature descriptors such as Histograms of Oriented Gradients (HOG), Histograms of Optical Flow (HOF), and motion boundary histogram. It takes the correlation between trajectories and target action into consideration. A saliency map based on optical flow has been introduced to highlight regions of foreground motion. Trajectory action relevance and frame action relevance are used as weights to identify discriminating trajectories and frames in a video during encoding.

Various features such as energy, sine, distinct body parts movements, and a 3D Cartesian view of smoothing gradients [5] have been used to extract full-body human silhouette features. Features extracted to represent human key posture points include rich 2D appearance, angular points, and multi-point autocorrelation. After the extraction of key points, a hierarchical classification and optimization model has been applied via ray optimization and a K-ary tree hashing algorithm.

The most common unsupervised learning platform, PCANet, is used for HAR [2]. Multiple Short-Time Motion Energy Image (ST-MEI) templates are computed to extract temporal features. The PCA-Net gives hierarchical motion features which are reduced using whitening PCA. As a classifier, Support Vector Machine (SVM) is used.

The method in [2] is extended by fusing spatial and temporal features learned from a PCANet in combination with Bag-of-Features (BoF) and Vector of Locally Aggregated Descriptors (VLAD) encoding schemes. Here, an encoding scheme is applied to represent the spatiotemporal PCANet features by feature fusion.

An effective algorithm has been developed using optical flow and wavelet transformation [20]. The two-fold transformation is done via the Gabor Wavelet Transform (GWT) and Ridgelet Transform (RT). The GWT produces a feature vector by calculating first-order statistics values of different scales and orientations of an input pose, which has robustness against translation, scaling, and rotation. The orientation-dependent shape characteristics of human action are computed using RT. The fusion of these features gives a robust unified algorithm.

An end-to-end two-stream attention based Long Short Term Memory (LSTM) network has been developed [4] for action recognition. This method focuses on the effective features of the original input images and pays different levels of attention to the outputs of each deep feature map. A deep feature correlation layer has been implemented to adjust the deep learning network parameters based on the correlation judgement.

A Spatio-Temporal Deep Residual Network with Hierarchical Attentions (STDRN-HA) has been designed for HAR [10]. In the first attention layer, the ResNet fully connected feature guides the Faster R-CNN feature to generate object-based attention for target objects. In the second attention layer, the O-attention further guides the ResNet convolution feature to yield the holistic attention (H-attention). In the third attention layer, the attention maps use the deep features to obtain the attention-enhancing features.

In [18], spatial and temporal features are extracted using a state-of-the-art 3D CNN from sampled snippets of a video. These features are then concatenated to form global representations,

which are then used to train a linear SVM for action classification. A. Nadeem developed a robust structure approach that discovers multidimensional features along with body part representations [11]. They used quadratic discriminant analysis with extracted features for human pose estimation and the maximum entropy Markov technique for classification.

Another approach has been developed which attempts to use motion tracking for human tracking as well as spatial feature extraction in a video sequence [6]. To have a strong feature vector for classification, Gaussian Mixture Model (GMM) and Kalman Filter (KF) methods are used to detect and extract the moving person, and Gated Recurrent Neural Networks are used to collect the features in each frame and predict human action.

A few recent research papers related to CBVR systems are discussed below.

A CBVR system has been designed for a large dataset [15]. This method uses vector motion-based signatures to describe visual content. It also uses keyframes for rapid retrieval. Instead of using low-level features, high-level features are considered in [12]. This method uses objects as features and compares the object with all other videos in the database.

An action template-based keyframe extraction method has been developed in [14]. This method also utilises keyframes for quick retrieval. For keyframe selection, informative regions are identified from each frame. Chen et al. have developed a supervised video hashing method based on deep 3DCNN for video retrieval [3]. This method converts the video features into binary codes using the hash function.

An unsupervised video retrieval method [24] has been developed that simultaneously models intra-frame and inter-frame contextual information for video representation with a graph topology that is constructed on top of pyramid regional feature maps. This method decomposes a frame into a pyramidal regional sub-graph and transforms a video into a regional graph. Thus, graph convolutional networks have been used to extract features that incorporate information from multiple types of context. A Modified Visual Geometry Group (MVGG\_16) [9] has been designed for the CBVR system. In this method, the video frame retrieval is performed by assigning an index to all video files in the database.

A multimodal CBVR [13] has been built that takes both the visual and audio information into account for retrieving relevant videos for the user. Two modules are employed by this method to deal with video and audio data. The video data is processed to detect the significant frame from shots and is achieved by the Lion Optimization Algorithm (LOA). The features are extracted from the visual data and, with respect to the audio data, Mean Hilbert Envelope Coefficients (MHEC) and Linear Predictive Cepstral Coefficients (LPCC) features are extracted. The extracted features are clustered by the Kernelized Fuzzy C Mean (KFCM) algorithm. Finally, the feature database is formed and is utilised in the query matching process during the testing phase.

### 3 Proposed HAR methodology

The proposed HAR system is designed with a network named STHAR. The proposed system architecture is shown in Fig. 1.

Actions can be recognised from a video using spatial and temporal features. Spatial features are extracted from frames, while temporal features are extracted from motion information. Instead of extracting spatial features from all frames, a few representative frames are selected from the input video. For selecting representative frames or keyframes, the Adaptive Group Of Pictures (AGOP) method [17] is used.

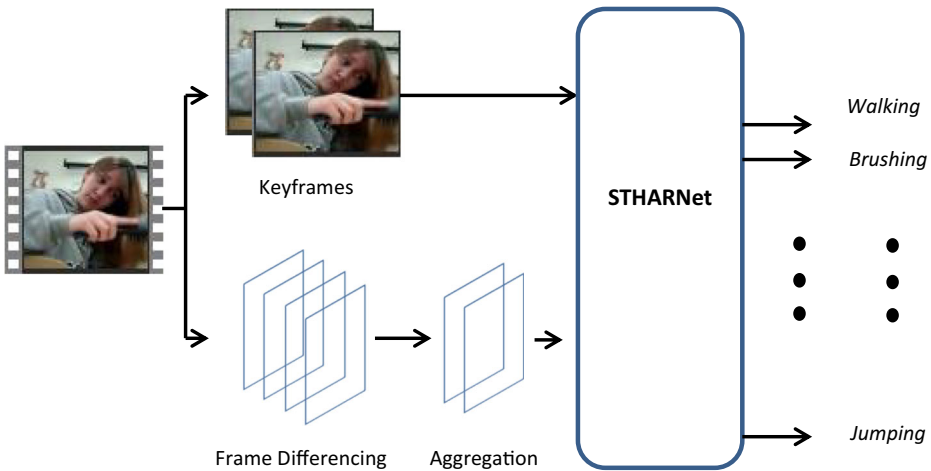


Fig. 1 Overall System Architecture of Proposed HAR Architecture

In the AGOP method, the video is divided into GOPs according to scene cut. From each GOP, the first and last frames are selected as keyframes. It should be noted that the last frame of a GOP is the first frame of the next GOP. The Motion Energy Image (MEI) is generated from each GOP to create motion information for that GOP. The temporal features are extracted from the MEI of each GOP. Both these features are fused and classified in the STHAR network.

### 3.1 Keyframe identification

The input video is divided into scenes (also named as GOP) based on the movement of objects in frames. The movement is identified by correlation component. In this work, Pearson Correlation Coefficient (PCC) is used which is given as

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

Where  $f$  and  $f_p$  are the frames for which the correlation is calculated.  $f^m$  and  $f_p^m$  are mean values of the frames  $f$  and  $f_p$  respectively. The absolute value of PCC ranges from 0 (no correlation) to 1 (perfect correlation). It is analysed that if the value of PCC lies below 0.8, there is no scene cut. If it goes above 0.8, then there is a cut. This value is chosen by few experiments.

A sample video is taken for this analysis whose scenes are identified manually. The number of frames for each scene obtained by the PCC based method is compared with manual scene identification for various values of PCC which is given in Table 1.

In Table 1, for Scene 1, the total number of frames is 10 using manual visualization. The PCC based scene identification method identifies only 1 correct frame for Scene 1. The remaining frames are misclassified to Scene 2, Scene 3 etc. When the threshold value increases, the number of correct frames in each scene matches the manual visualization. When the threshold is 0.9, the number of correctly classified frames reduces. Here, the frames in the subsequent scenes are misclassified to the current scene (i.e. frames in the Scene 2 are misclassified as Scene 1). Hence, it is clear that the number of frames in every scene is correctly identified when keyframe threshold is set to 0.8

**Table 1** Comparison of Manual and PCC based Scene Identification

Scenes	Manual (No. of frames)	PCC Based (No. of Correct Frames) for Threshold						
		0.3	0.4	0.5	0.6	0.7	0.8	0.9
Scene 1	10	1	2	4	6	8	10	10
Scene 2	12	1	2	5	7	10	12	10
Scene 3	9	1	2	3	5	7	9	6
Scene 4	15	1	3	5	8	12	14	11

After scene cut, the first frame in each GOP is selected as keyframe. Spatial features are extracted from these keyframes.

### 3.2 Motion Energy Image

From each GOP created in the previous subsection, a Motion Energy Image is generated from which temporal features are extracted. MEI is a binary cumulative motion image which is obtained using either background subtraction or frame differencing. Frame differencing is a simple and efficient solution to a dynamic background problem. In this work, multiple GOP-MEI templates are computed using simple frame differencing to preserve local motion information and to efficiently represent human movements that occurred in the image sequence.

Assume that  $f(x, y, t)$  and  $f(x, y, t + 1)$  represent two consecutive frames in the input video data acted at time  $t$  and  $t + 1$ , respectively. The absolute frame difference  $diff(x, y, t)$  of these two frames can be obtained as:

$$diff(x, y, t) = |f(x, y, t) - f(x, y, t + 1)| \quad (2)$$

The  $diff(x, y, t)$  is binarized as  $D(x, y, t)$  by Otsu thresholding. For each GOP, MEI is created as

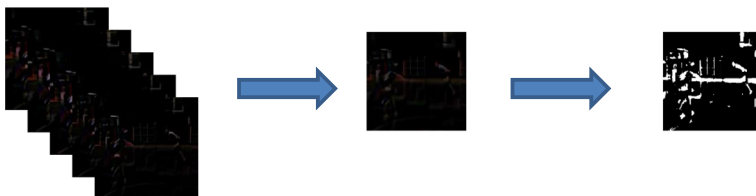
$$E_g(x, y, t) = \sum_{i=1}^n D(x, y, t - i) \quad (3)$$

Where  $E_g(x, y, t)$  is the Motion Energy Image at time  $t$  for  $n$  number of frames in a GOP. Figure 2 shows the frame differencing frames for a GOP and the obtained MEI.

The whole action video ( $V$ ) can be expressed as follows:

$$V = \{E_g\}_{g=1}^G \quad (4)$$

Where  $G$  is the number of GOP.



**Fig. 2** Motion Energy Image

### 3.3 STHAR network

A novel architecture is designed to extract spatial and temporal features from keyframes and MEI respectively, which is named as STHAR Net. The architecture of the STHAR network is shown in Fig. 3. This network is composed of spatial stream ConvNets and temporal stream ConvNets. The spatial stream ConvNets work on keyframes and temporal stream ConvNets work on MEI. In this work, the Inception with Batch Normalization (BN-Inception) network architecture is used as a building block, as this network has good balance between efficiency and accuracy. The dropout ratio is set to 0.7 for spatial stream ConvNet and 0.8 for temporal stream ConvNet. The framework of ConvNet consists of 3 layers. Every layer consists of a batch normalization, a convolution layer, and a Rectified Linear Unit (ReLU) activation function.

### 3.4 Content based video retrieval

The conventional content based video retrieval system consists of training and testing phase. In training phase, the training videos are divided into video shots from which the features are extracted and stored in a database. In testing phase, the features of the query video shot are extracted. These features are matched with all the features in the features database. Few top relevant video shots are retrieved as output. In the proposed CBVR system, the above discussed STHAR is integrated.

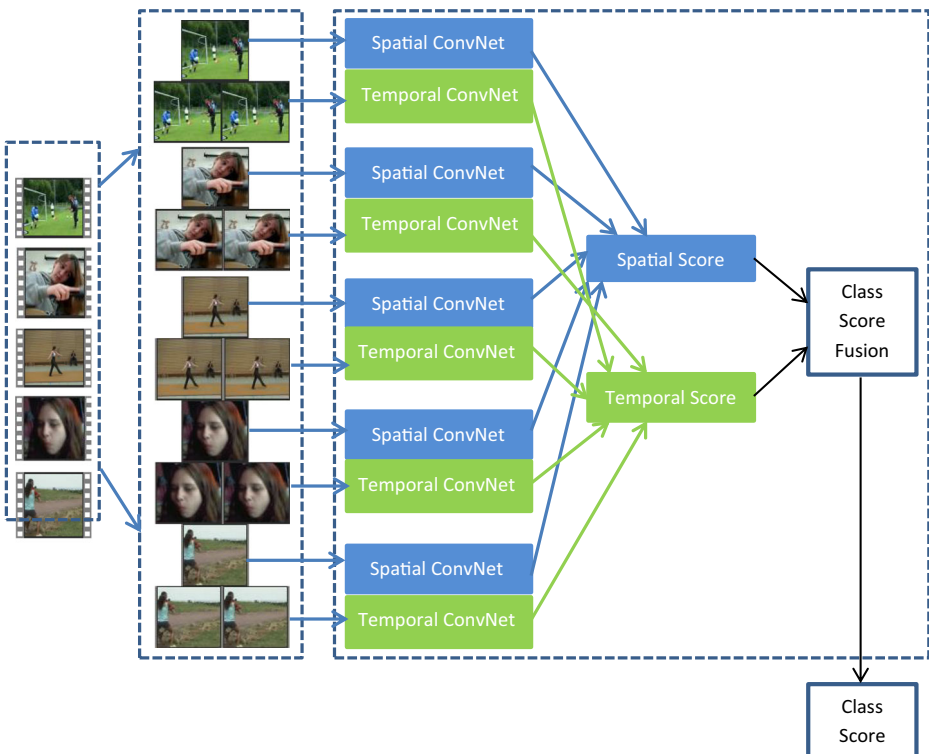


Fig. 3 Proposed STHAR Network

Figure 4 illustrates the proposed STHAR-CBVR system. In STHAR system, the features are used to obtain the score of each action. In CBVR system, the features are fused and stored in a dataset. We concatenate the spatial and temporal features using a 1D convolution. For query image, spatio-temporal features are extracted and compared with the feature dataset using similarity measure. In this work, Euclidean distance is used as similarity measure.

The steps of STHAR and CBVR integration are given below:

**Step 1:** Divide the videos into scenes using keyframe extraction method.

**Step 2:** The spatial and temporal features are extracted separately from keyframes and intermediate frames respectively.

**Step 3:** The spatial and temporal scores are generated and fused to obtain the class score for the video.

**Steps 4:** The above steps form STHAR system.

**Step 5:** For a CBVR system, the spatial and temporal features extracted in Step 2 are fused and stored in features dataset.

**Step 6:** For a given query scene, the spatial and temporal features are fused and matched with the features datasets using similarity measure.

**Step 7:** All the similar videos that matches the feature are retrieved from the database.

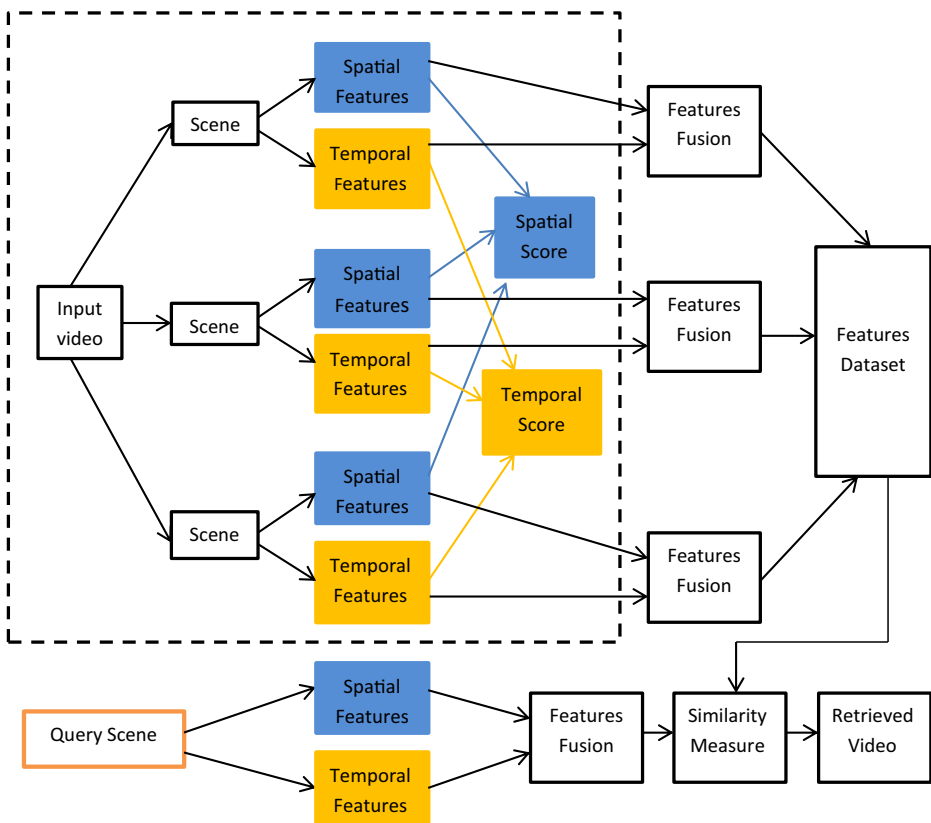


Fig. 4 Proposed STHAR-CBVR System



The algorithm for Proposed STHAR-CBVR system is shown in Algorithm 1.

---

**Algorithm 1** Proposed STHAR-CBVR System

---

**Input:** Video Dataset  $D$ , Query video frames  $Q$

**Output:** Retrieved video frames  $R_f$ , Score

**Steps:**

//STHAR

1. For each video  $V$  in  $D$

    //Calculate Keyframes

    1.1  $K_f = f_i$

    1.2 For each frame  $f_p$  and  $f_{p+1}$  in  $V$

$$1.2.1 \ PCC = \frac{\sum_{i=1}^M \sum_{j=1}^N (f_p(i,j) - f_p^m)(f_{p+1}(i,j) - f_{p+1}^m)}{\sqrt{\sum_{i=1}^M \sum_{j=1}^N (f_p(i,j) - f_p^m)^2 (f_{p+1}(i,j) - f_{p+1}^m)^2}}$$

    1.2.2 If  $PCC > 0.8$

        Add  $f_{p+1}$  to  $K_f$

    1.2.3 End

    1.3 End

2. End

3. Extract  $(f_s, f_t)$  are from keyframes  $K_f$  and intermediate frames using Spatial and Temporal ConvNet respectively.

4. The above step produces the scores  $(Sc_{Spatial}, Sc_{Temporal})$ .

5. Score =  $\text{mean}(Sc_{Spatial}, Sc_{Temporal})$

//CBVR

6. Concatenate  $(f_s, f_t)$  obtained in Step 3 in a dataset  $D_f$ .

7. For a given query scene  $Q$ , obtain  $(f_s, f_t)$

    //Match the features

$$8. \ ED = \sum_{i=1}^n \sqrt{\sum_{i=1}^n ((f_s, f_t)_{pi} - (f_s, f_t)_{qi})^2}$$

9. All the similar videos that match the feature are retrieved from the database.

---

## 4 Experimental results

This section discusses and analyses the proposed method by exposing the obtained results. It also gives a brief description of the datasets used, results comparison with other methods and ablation study.

## 4.1 Dataset description

The proposed method is tested on three publicly available datasets: HMDB51 [8], KTH [16] and UCF Sports action [19] dataset. The details of the dataset are given in Table 2. Among the three datasets, HMDB51 has complex actions. The HMDB51 dataset consists of 6766 videos of 51 actions whereas KTH and UCF Sports dataset consists of 6 and 10 actions respectively. The dataset is divided into 80% training and 20% testing. Each dataset is of different sizes.

## 4.2 Implementation details

The mini-batch Stochastic Gradient Descent (SGD) algorithm is used to learn the network parameters. Two SGD optimizers are used in this work (SGD<sub>1</sub> and SGD<sub>2</sub>).

$$w_{k+1} = w_k - (\lambda_a \cdot \eta + \lambda_b \cdot \eta_b) \cdot (\lambda_a \cdot d_a + \lambda_b \cdot d_b) \quad (5)$$

Where  $\lambda_a$  is a scalar for SGD<sub>1</sub> and  $\lambda_b$  is another scalar for SGD<sub>2</sub> for balancing the two contributions of the two optimizers.  $\eta$  is the learning rate of the created optimizer.  $\eta_b$  is the learning rate of SGD<sub>2</sub> optimizer.  $d_a$  and  $d_b$  are the two increments of SGD<sub>1</sub> and SGD<sub>2</sub> using their weights in each iterations respectively. We changed the optimizer lambda dynamically from  $\lambda_a=1, \lambda_b=0$  and  $\lambda_a=0, \lambda_b=1$ .  $\eta$  is set to 1 which is multiplied by 0.1 at every 100 epochs.

Table 3 shows the hyper parameters for both the networks. For data augmentation, horizontal flipping, corner cropping and scale jittering are used. The proposed method is tested on Intel Core i7 2.8 GHz machine with a GPU NVIDIA GeForce GTX 1050.

## 4.3 Performance metrics

The proposed method includes two major research works: HAR and CBVR. The performance of the HAR system is measured using accuracy. Similarly, the performance of the CBVR system is measured using precision, recall, F-score and sensitivity. All the performance metrics used in this work use 4 major values: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The formula for all the metrics is shown in Table 4.

## 4.4 Results

As two major phases are used in this research, both phases are analyzed in terms of performance. The visual and quantitative results are analysed in this section. The accuracy of CBVR is based on the relevance of the output set of images or videos to a particular query. A user defines the level of quality for the evaluation of CBVR. The accuracy of these systems is purely subjective, which requires some human intervention during evaluation.

**Table 2** Properties of Datasets

Property	HMDB51	KTH	UCF Sports
Number of Action Classes	51	6	10
Number of Video Clips	6766	600	150
Resolution	320×240	160×120	720×480
Frame Rate	30 fps	25 fps	-

**Table 3** Experimental Setup of Spatial and Temporal ConvNet

Hyper Parameter	Mini-batch	Momentum	Learning Rate	Dropout
Spatial ConvNet	256	0.9	0.001	0.7
Temporal ConvNet	256	0.9	0.005	0.8









**Table 4** Performance Metrics used in HAR and CBVR

Metrics	Formula
Precision	$Precision = \frac{TP}{TP+FP}$
Recall	$Recall = \frac{TP}{TP+FN}$
F-Score	$F_m = 2 \frac{Precision * Recall}{(Precision+Recall)}$
Sensitivity	$Specificity = \frac{TN}{(FP+TN)}$
Accuracy	$Acc = \frac{TP+TN}{TP+TN+FP+FN} \times 100$

A subject is generally shown a query and asked to rank the resulting image or video segments on a scale. For example, in a video database the user might be asked to rate the quality of selection on a scale ranging from “high relevance” (Rank – 1) to “low relevance” (Rank – 10). We consider the top 10 relevant results of the given query image.

Accuracy is obtained using top 10 outputs of the retrieved results and the top 10 results obtained by subjective evaluation. The visual results obtained by STHAR-CBVR system are shown in Table 5. Only 5 frames from the retrieved results are shown in the table. All the frames in the table correspond to HMDB51 dataset. As this research is the integration of HAR and CBVR, the class name is also shown in Table 5.

**Table 5** Results obtained by STHAR-CBVR system in HMDB51 dataset

Keyframe in Query Video	Class	Retrieved Results (Shown only 5 frames)
	Chew	
	Brush hair	
	Cart wheel	
	Catch	

**Table 6** Accuracy Obtained by the Proposed STHAR system for all the Compared Datasets

Method	Measure	HMDB51	KTH	UCF Sports
Proposed STHAR	Accuracy	89.9	98.7	99
Proposed STHAR-CBVR	Precision	73	74.3	74.9
	Recall	72.8	74.2	74.5
	F-Score	72.6	73	74.69
	Sensitivity	71	74	75
	Accuracy	73.25	74.6	75.3

The difference between HAR and CBVR is that the HAR shows the name of the action class; whereas CBVR shows the video based on the query video. The proposed STHAR-CBVR system also shows some other videos related to the content. But only few frames are shown in Table 5.

From Table 5, it is observed that the retrieved outputs are very similar to the query frame. The classes thus shown in Table 5 are the output from proposed HAR system. Thus, the proposed STHAR-CBVR is able to identify the class and obtain similar videos from dataset for a given query scene. The accuracy obtained by the proposed STHAR system is shown in Table 6.

From Table 6, it is observed that the accuracy obtained by the proposed STHAR system is greater than 90% for all the datasets. It is also studied that the precision, recall, F-score, sensitivity and accuracy of proposed STHAR-CBVR system are in the range of 71–75%. The accuracy of the retrieval system is much lower than recognition system. The confusion matrices of the proposed STHAR system for KTH and UCF Sports Action dataset are given in Fig. 5.

In KTH dataset, we reached almost 99% accuracy for 4 classes. In UCF Sports action dataset, we reached maximum accuracy for Diving-side, Golf-swing and Swing-Bench classes. The proposed STHAR system achieves 99% accuracy on 4 classes of UCF sports dataset.

#### 4.5 Comparison of proposed methods with recent methods

The proposed STHAR and STHAR-CBVR methods are separately compared with recent methods [1, 2, 4–7, 10, 11, 14, 15, 18, 20–23] that are discussed in Section 2. Table 7 shows the comparison of proposed STHAR method with recent methods for the tested datasets.

From Table 7, it is studied that there is a 0.7% increase in accuracy in HMDB51 dataset when compared to Pseudo 2D stick model [5]. Similarly, there is a 0.4% increase in accuracy in KTH dataset. But in UCF Sports dataset, accuracy comes down by 0.2%. Table 8 shows the comparison of retrieval system with other retrieval systems.

From Table 8, it is inferred that the proposed method achieves higher accuracies than recent methods on HMDB51 and KTH dataset.

#### 4.6 Ablation study

The proposed method is evaluated by two different experiments. The spatial and temporal features are studied by using them separately. The pre-trained network models are also varied to measure the accuracy of the proposed methods. Table 9 shows the first study and Table 10 shows the second study.

Confusion Matrix

True Label	Boxing	0.98	0.01	0.01	0.00	0.00	0.00
	Handclapping	0.00	0.99	0.00	0.01	0.00	0.00
	Handwaving	0.01	0.00	0.99	0.00	0.00	0.00
	Jogging	0.00	0.00	0.00	0.99	0.00	0.01
	Running	0.00	0.00	0.00	0.00	0.99	0.01
	Walking	0.00	0.00	0.00	0.01	0.01	0.98
		Predicted Label					

(a)

Confusion Matrix

True Label	Diving Side	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Golf-swing	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Kicking front	0.00	0.00	0.99	0.00	0.00	0.01	0.00	0.00	0.00
	Lifting	0.00	0.00	0.00	0.99	0.00	0.00	0.01	0.00	0.00
	Riding Horse	0.00	0.01	0.00	0.00	0.99	0.00	0.00	0.00	0.00
	Run Side	0.00	0.00	0.01	0.00	0.00	0.99	0.00	0.00	0.00
	Skate-Boarding Front	0.00	0.00	0.00	0.01	0.00	0.01	0.98	0.00	0.00
	Swing Bench	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
	Swing Side Angle	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.98
	Walk Front	0.00	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.98
		Predicted Label								

(b)

Fig. 5 Confusion Matrix Obtained by the Proposed STHAR system for (a) KTH dataset (b) UCF Sports Action Dataset

**Table 7** Comparison of Proposed STHAR Method with Recent Methods

Dataset/Method	HMDB51	KTH	UCF Sports
Khan et al. [7]	81.4	98.3	99.2
Wang et al. [21]	75.2	-	-
Yudistra et al. [23]	69	-	-
Yi et al. [22]	74.6	97.83	-
Pseudo 2D Stick Model [5]	89.21	-	-
STMEI-PCANet [2]	-	-	86.7
ST-VLAD-PCANet [1]	-	93.33	90
Vishwakarma [20]	-	96.6	-
Dai et al. [4]	-	-	96.9
Li et al. [10]	70.61	-	-
Torpey et al. [18]	62.8	-	-
Jaouedi et al. [11]	89.09	-	90.91
Jaouedi et al. [6]	-	-	89.01
Proposed STHAR Method	89.9	98.7	99

From the above table, it is studied that the two stream network achieves higher accuracy than individual streams on all the datasets. The accuracy is very lesser when only spatial ConvNets are used. The most popular Googlenet and VGG16 are used for spatial and temporal ConvNets and the average results are shown in Table 10.

From Table 10, it is found that BN-Inception network model achieves higher accuracy than GoogleNet and VGG16 models.

**Table 8** Comparison of Proposed STHAR-CBVR Method with Other Methods

Method	HMDB51	KTH
Saudi et al. [15]	72.37	-
Savran et al. [14]	-	71.13
Proposed Method	73.25	74.6

**Table 9** Accuracy Study of Proposed STHAR-CBVR systems for various ConvNets

Dataset	Spatial ConvNets	Temporal ConvNets	Two Stream
HMDB51	71	72.4	73.25
KTH	71.2	72	74.6
UCF Sports	73.7	74.8	75.3

**Table 10** Average Accuracy Study of the Proposed HAR-CBVR system for various Pre-trained Networks

Dataset	HMDB51	KTH	UCF Sports
GoogleNet	72.6	72.23	74.45
VGG16	73	73.4	74.64
BN-Inception	73.25	74.6	75.3

## 5 Conclusion

Human action recognition and content-based video retrieval systems occupy huge areas in research field. This paper combines both methods by introducing a novel STHARNet in the HAR system. The proposed system is integrated with CBVR system and satisfies both systems. It is tested on major datasets such as KTH, HMDB51 and UCF Sports action datasets and proved its efficacy over other methods. The proposed STHAR and STHAR-CBVR systems achieve accuracy on an average of 98% and 74% for all the datasets respectively. The proposed system can be further extended to use some other pre-trained networks and can be tested on more complicated datasets.

**Data availability** The dataset used in this research are publicly available from the following websites:

HMDB51: <https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/>.

UCF101: <https://www.crcv.ucf.edu/data/UCF101.php>.

UCF Sports: [http://crcv.ucf.edu/data/UCF Sports Action.php](http://crcv.ucf.edu/data/UCF%20Sports%20Action.php)

## Declarations

**Conflict of interest** This work entitled “STHARNet: Spatio-Temporal Human Action Recognition Network in Content Based Video Retrieval” is not submitted anywhere else. There is no conflict of interest from authors.

## References

1. Abdelbaky A, Aly S (2021) Two-stream spatiotemporal feature fusion for human action recognition. *Visual Comput* 37(7):1821–1835
2. Ahmed A, Aly S (2020) Human action recognition using short-time motion energy template images and pcanet features. *Neural Comput Appl*:1–14
3. Chen H, Hu C, Lee F, Lin C, Yao W, Chen L, Chen Q (2021) A supervised video hashing method based on a deep 3D convolutional neural network for large-scale video retrieval. *Sensors* 21(9):3094
4. Dai C, Liu X, Lai J (2020) Human action recognition using two-stream attention based LSTM networks. *Appl Soft Comput* 86:105820
5. Jalal A, Akhtar I, Kim K (2020) Human posture estimation and sustainable events classification via pseudo-2D stick model and K-ary tree hashing. *Sustainability* 12(23):9814
6. Jaouedi N, Boujnah N, Bouhlelc MS (2020) A new hybrid deep learning model for human action recognition. *J King Saud Univ Comput Inf Sci* 32:447–453
7. Khan MA, Zhang YD, Khan SA, Attique M, Rehman A, Seo S (2021) A resource conscious human action recognition framework using 26-layered deep convolutional neural network. *Multimed Tools Appl* 80(28): 35827–35849
8. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: a large video database for human motion recognition. In: 2011 International Conference on Computer Vision. IEEE, pp 2556–2563
9. Kumar BS, Seetharaman K (2022) Content based video retrieval using deep learning feature extraction by modified VGG\_16. *J Ambient Intell Humaniz Comput*:1–13
10. Li Y, Liu C, Ji Y, Gong S, Xu H (2020) Spatio-temporal deep residual network with hierarchical attentions for video event recognition. *ACM Trans MCCA* 16:1–21
11. Nadeem A, Jalal A, Kim K (2020) Accurate physical activity recognition using multidimensional features and Markov model for smart health fitness. *Symmetry* 12:1766
12. Pinge A, Gaonkar MN (2021) A novel video retrieval method based on object detection using deep learning. In: *Computational vision and bio-inspired computing*. Springer, Singapore, pp 483–495

13. Prathiba T, Kumari RSSP (2021) Content based video retrieval system based on multimodal feature grouping by KFCM clustering algorithm to promote human–computer interaction. *J Ambient Intell Humaniz Comput* 12(6):6215–6229
14. Savran Kızıltepe R, Gan JQ, Escobar JJ (2021) A novel keyframe extraction method for video classification using deep neural networks. *Neural Comput Appl*:1–12
15. Saoudi EM, Jai-Andaloussi S (2021) A distributed content-based video retrieval system for large datasets. *J Big Data* 8(1):1–26
16. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. IEEE, vol 3, pp 32–36
17. Sowmyayani S, Rani J, Arockia P (2014) Adaptive GOP structure to H. 264/AVC based on Scene change.&nbsp;&nbsp;&nbsp;ICTACT J Image Video Process 5(1)
18. Torpey D, Celik T (2020) Human action recognition using local two-stream convolution neural network features and support vector machines. arXiv arXiv:2002.09423. Available online: <https://arxiv.org/abs/2002.09423> . Accessed on 19 Feb 2020
19. UCF Sports Website : [http://crcv.ucf.edu/data/UCF\\_Sports\\_Action.php](http://crcv.ucf.edu/data/UCF_Sports_Action.php)
20. Vishwakarma DK (2020) A two-fold transformation model for human action recognition using decisive pose. *Cogn Syst Res* 61:1–13
21. Wang J, Peng X, Qiao Y (2020) Cascade multi-head attention networks for action recognition. *Comput Vis Image Underst* 102898
22. Yi Y, Li A, Zhou X (2020) Human action recognition based on action relevance weighted encoding. *Signal Process Image Commun* 80:115640
23. Yudistira N, Kurita T (2020) Correlation net: spatiotemporal multimodal deep learning for action recognition. *Signal Process Image Commun* 82:115731
24. Zhao G, Zhang M, Li Y, Liu J, Zhang B, Wen JR (2021) Pyramid regional graph representation learning for content-based video retrieval. *Inf Process Manag* 58(3):102488

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.