Soft Computing
https://doi.org/10.1007/s00500-021-06149-7

APPLICATION OF SOFT COMPUTING

Human action recognition using a hybrid deep learning heuristic

Samarendra Chandan Bindu Dash[1] · Soumya Ranjan Mishra[2] · K. Srujan Raju[3] · L. V. Narasimha Prasad[4]

Accepted: 13 August 2021

Abstract

Human action recognition in the surveillance video is currently one of the challenging research topics. Most of the works in this area are based on either building classifiers on sophisticated handcrafted features or designing deep learning-based convolutional neural networks (CNNs), which directly act on raw inputs and extract meaningful information from the video. To capture the motion information between adjacent frames, 3D CNN extracts features in temporal dimension along with spatial dimension. Even though this technique is very effective in human action recognition but limited to very few fixed frames, all human actions are not limited to a fixed number of frames; they may span several frames. If we increase the size of the input window in CNN, handling all trainable parameters in the network will be very complicated. Hence, it is advisable to encode high-level motion features from different sources to the CNN model. This paper proposed a novel framework to extract handcrafted high-level motion features and in-depth features by CNN in parallel to recognize human action. SIFT is used as handcrafted feature to encode high-level motion features from the maximum number of input video frames. The combination of deep and handcrafted features preserves more extended temporal information from entire video frames present in action video with minimal computational power. Finally, we pass the extracted SIFT into the dense layer and concatenate it with a fully connected layer of CNN for classification. We evaluate the proposed combined CNN framework against regular 3D CNN and traditional handcrafted features like optical flow with SVM, SIFT with SVM on UCF, and KTH human action dataset. We achieve better performance in terms of computational cost and processing time in the proposed CNN framework compared to the other three methods.

Keywords Human action recognition · 3D Convolutional Neural network · Deep Neural Network · Optical flow · SIFT · Motion features extraction

1 Introduction

Human action recognition in a real-world environment is highly applicable in various domains like video surveillance systems, crowd behavior analysis, human-robot interaction. However, recognizing the accurate action in real-world video is a challenging task due to many factors like variation in viewpoint, scale, cluttered background and many more (Feichtenhofer et al. 2017; Junejo et al. 2011; Le et al. 2011; Wang and Mori 2011). Few authors had taken certain assumptions on these factors while taking video and had shown comparatively better results on this problem (Jhuang et al. 2007; Ramezani and Yaghmaee 2016). Generally, human action recognition is a two steps process. In the first step, different types of features need to be extracted from the raw video frame. Then in the second step, the classifier needs to be trained on extracted features to categorize different action classes. Feature extraction is one of the very crucial parts of video and image analysis. Generally, two types of feature

✉ K. Srujan Raju
ksrujanraju@gmail.com

Samarendra Chandan Bindu Dash
samarendra109@gmail.com

Soumya Ranjan Mishra
soumyaranjanmishra.in@gmail.com

L. V. Narasimha Prasad
lvnprasad@iare.ac.in

[1] Informatica Pvt. Ltd, Bengaluru, India

[2] Department of Computer Science and Engineering, ANITS, Visakhapatnam, India

[3] Department of Computer Science and Engineering, CMR Technical Campus, Hyderabad, India

[4] Department of Computer Science and Engineering, Institute of Aeronautical Engineering, Hyderabad, India

Published online: 28 August 2021

extraction methods are widely used for video representation; the first is handcrafted-based feature descriptors, and the second is deep learning-based descriptors. CNN (Convolutional Neural Network) is a deep learning-based descriptor where convolution and pooling operation is applied in different layers on the raw input image and produces a fully connected layer. After several layer of trainable filters and pooling, it generates hierarchy of complex features, which is very efficient on visual object recognition task (Donahue et al. 2017; Simonyan and Zisserman 2014a; Tran et al. 2015; Yu et al. 2009). There are few works on human action recognition by 2D CNN model (Ning et al. 2005) by representing single frame architecture, where the feature vector is extracted for each frame. This 2D CNN architecture can be used to train on a large-scale human action dataset, but the major problem in 2D CNN is that it does not capture and model the temporal information between different frames. Human action videos consist of multiple frames; therefore, 3D CNN is more effective to model both spatial and temporal information in a short period (Ji et al. 2013). However, 3D CNN shows tremendous success for human action classification from video clips but is limited to a fixed number of input frames. In real-time, human action may be performed throughout the entire duration of the video. We need a huge computational cost to run 3D CNN on the complete sequence of 2D frames. It takes around 3 to 4 days on UCF101 and nearly two months on sports-1M to train a 3DConvNet with very extensive architecture (Tran et al. 2017). Handcrafted-based SIFT descriptors and bag of visual word (BOW) models achieve state-of-the-art performance for human action recognition and better compromise between computational complexity and recognition rate. In our previous work, scale-invariant feature transform (SIFT) features were used in Bezier cohort fusion in doubling states for human identity recognition (Garain et al. 2019). Considering both handcrafted and in-depth CNN features, a novel combined framework is introduced to capture motion information from the entire video clip (all sequence of frames) with limited computational cost and complexity. By this combination, both high and low-level spatiotemporal information can be preserved from the entire video clip with minimum computational cost.

Inspired by both manual handcrafted features extraction and bio-inspired based convolution operation, in this work, we have combined both in different layers of 3D CNN. In the first layer, we have extracted features by applying a hardwired kernel to the first seven frames of the input video and generates multiple channels of information and then applied multiple layers of convolution and pooling in both spatial and temporal dimensions to all the channels separately. Along with convolution, we have extracted key point-based SIFT descriptors from raw input video frames (all the frames present in action video) to model temporal information with minimal computational cost and time. Finally, we pass the extracted SIFT into the dense layer and concatenate it with a fully connected layer of CNN for classification.

We evaluate the proposed combined CNN framework on UCF and KTH human action dataset. We have taken the first seven frames for CNN and the entire video clip (all the frames) for SIFT in parallel for feature extraction. The concatenation of two feature vectors is used for classification. Our experiments show that the proposed combined CNN framework outperforms better than state-of-the-art human action recognition systems with very less computational cost and minimal processing time. The major contribution of this proposed model is summarized as below:

- First, handcrafted feature SIFT is extracted from the raw input video clip(all frames) followed by PCA to form a fully connected fixed-size vector.
- In parallel with SIFT, we extracted features like gradient images, optical flow, and gray channel for the initial seven cropped input frames for convolution operation. The Gradient of an image can be obtained by applying the Sobel operation in the X and Y direction; similarly, optical flow is also computed in the X and Y direction between two consecutive frames. Gray code channel is also taken for all frames to generate different channels of information.
- We then apply multiple layers of convolution and polling in both spatial and temporal domain with a fixed kernel size to all different channels separately and generates a fully connected layer.
- Finally, we pass the extracted SIFT into the dense layer and concatenate with a fully connected layer of CNN followed by a dropout with sigmoid for final classification.
- We evaluate our proposed model on UCF and KTH human action dataset and observed interesting results as compared with other state-of-the-art human action recognition models.

2 Related work

In this literature, works on human action classification such as traditional handcrafted spatio-temporal features representation and deep learning-based approach followed by related combination strategies are described. Predominantly many human action recognition models consist of two primary modules, action representation and action classification. Converting feature vectors from an action video is generally referred to as action representation and conclude an action label from extracted feature vector is termed as action classification. Modeling the human action from different features is a traditional machine learning technique, where different motion information of input video is extracted and trained to the classifier. In this section, we will discuss the various handcrafted feature and deep learning-based features for human action recognition relevant to our proposed method.

2.1 Handcrafted feature representation

Handcrafted features are designed to encode the movement of the human body, and spatiotemporal changes in action video (Patel et al. 2018; Choutas et al. 2018) including STIP-based method, spatiotemporal volume-based, and trajectory of skeleton joints methods. All these action features are used in traditional machine learning techniques like SVM, boosting, and probability map models to recognize human action. The earliest approach to human action recognition is proposed by Bobick and Davis (2001) using motion energy image(MEI) and motion history image(MHI). The main idea behind this model is to design an action template and perform template matching. In Zhang et al. (2008), polar coordinates are used to divide the human body in MHI, and SIFT-based motion descriptor is used to represent human action. Histogram of gradient (HOG) method is extended to space-time dimension and proposed 3DHOG to describe motion in video (Klaser et al. 2008). STIP-based techniques are the most widely used method for human action detection, in which the interest points are tracked in spatiotemporal dimension (Dawn and Shaikh 2016; Laptev 2003). 3D-Harris spatiotemporal feature points and some advanced techniques are proposed in space-time domain for effective human motion computation(Laptev 2005; Chakraborty et al. 2012). In Nguyen et al. (2014) proposed key region extraction technique by using spatiotemporal attention mechanism to design action features and visual dictionary. Again in Peng et al. (2016), a hybrid super vector human action representation method is proposed by modifying local spatiotemporal features and visual dictionary construction. To extract the key region of action video, 3D-Harris spatiotemporal information and 3D scale-invariance feature are combined, and visual word histograms are used to model human action (Nazir et al. 2018). Earlier, STIP-based human action recognition method attracted the attention of many researchers. The method mentioned above shows excellent results in the static background, but in the case of camera motion, it generates many unnecessary key points, which creates problems in recognizing object motion. A features weighting framework based on the movement of different body parts has also been proposed in Mishra et al. (2020) for efficient human action recognition.

Generally, the human joints (joint skeleton features) are invariant to camera motion and the appearance of an object. It provides the position of human joints accurately from a different viewpoint(front, back, side). Recently, joints in the human skeleton are used to represent human action. Improved dense trajectories (IDT) is proposed (Wang and Schmid 2013) to address human action from skeleton joints. The displacement of the extracted human joints(feature points) is calculated using the optical flow approach. Few researchers have improved the skeleton joint-based approaches by various ways (Wang et al. 2016; Peng et al. 2014). Local motion trajectories are analyzed by using split clustering (Gaidon et al. 2014) and used to represent motion level in human action. Stacked fisher vector also used in IDT feature to analyzed human action. The main advantage of this trajectory-based method is viewpoint and appearance invariant. However, we need an accurate human skeleton joints detector of the human body for this to model the motion of those points for effective action recognition.

2.2 Deep feature extraction

Now a days, deep neural networks combined both motion representation and classification components and represented a complete framework for action recognition problems, which enhances the classification accuracy (Núñez et al. 2018; Kong et al. 2017). Based on convolution operation, different types of 3D CNN architecture can be designed. The selection of architecture is entirely problem-dependent and specific to a particular application. Multi-stream CNN is proposed in Tu et al. (2018) to incorporate multiple features like full frame, human body, and motion-salient body part regions together in a single network. Region sequence-based CNN is proposed in Ma et al. (2018) for human action recognition in small-scale image regions. Here, pooling layer information is encoded for more spatial information. Very recently, a capsule network (capsnet) is introduced for the human action recognition problem (Algamdi et al. 2019). This capsnet replaces neurons with vectors which can retain spatial features relationship between frames. As compared to neurons-based CNN, this vector-based capsnet has shown better performance with the same complexity on KTH and UCF datasets. By mimicking the neural processing technique in the visual cortex of the human brain, a sparse-based neural response method is proposed in Li et al. (2014) for image classification. Biological modeling of the human vision system is proposed by Deng et al. (2018), in which very low-level visual processing is concerned. In this model, they simulated the properties of the retina and lateral geniculate nucleus (LGN) in the early stages. A combined RGB and skeleton data is used in CNN for human activities recognition (Khaire et al. 2018). Here, multiple vision cue is combined for competitive results. In Wang et al. (2019), traditional handcrafted motion features are trained by the CNN-based hallucination step. In this model, they combined dense trajectory of video frames with the output of CNN. In Zhu et al. (2016), a novel regularization scheme is proposed on human skeleton joint to learn co-occurrence features using LSTM network. A twofold approach has been proposed for the frame extraction and action recognition from video using a typical convolutional neural network (Mishra et al. 2020).

3 Proposed model for human action classification

3.1 Overview of the framework

Figure 1 shows an overview of our model. Our proposed method jointly trained both deep and handcrafted features from input videos. We have used two feature extraction methods: deep features (using CNN) and handcrafted features (using SIFT). Deep features CNN deals with a fixed number of frames to avoid complexity in the network. SIFT is calculated from all the 2D frames present in action video to extract longer temporal features. As we are extracting SIFT from the maximum number of frames, it increases the feature vector's dimension. To address this, we have used PCA on extracted SIFT to select fewer components. These two feature extraction methods are explained in detail in Sects. 3.2 and 3.3. Using these methods, we obtained two feature vectors that capture both spatial and temporal motion information from the entire input video clip. Finally, two feature vectors are concatenated and classified at the classification layer of CNN.

3.2 Handcrafted features SIFT

After various research and experiment on image and video analysis for human action recognition problems, one can conclude that local patches sampled around detected interest points preserve the most discriminative and adequate information on human actions. The local patches represent the complete action video without any further preprocessing like background subtraction or object tracking. This paper has used 3D SIFT as additional handcrafted features since it is robust and invariant to rotation and noise. The original 2D SIFT (Lowe 2004) is extended to 3D video volume where the third dimension is time, to capture space-time information. Especially for consecutive frames, motion edge history image(MEHI) proposed in Bobick and Davis (2001) preserve complete motion information with sufficient processing capability. Using this idea, first, we extracted moving points and tracked them throughout the entire video frames. Then, we extract SIFT descriptor around every detected point to describe the spatial information on frames. The displacement of those extracted points is described with trajectories. Inspired by the work in Kawai et al. (2010), moving points detection and their displacement computation is done by two widely used techniques, background subtraction, and Lucas–Kanade optical flow, respectively. Finally, a bag of features model is used to generate a histogram for every input action video. The complete process of feature extraction from input video is shown in Fig. 2 and explained below:

1. First, each current video frame is subtracted from the reference frame to obtained motion region or foreground.
2. After getting motion region, salient points are detected from these regions using Shi Tomasi corner detector (Tomasi and Detection 1991) to generate moving points.
3. Then, we extract SIFT descriptor around every detected point in both spatial and temporal domain. Spatial domain information computes appearance around the points, and temporal domain information represents the change of the feature points along the time axis in the 3D space-time volume.
4. Finally, a fully connected fixed-size vector is formed by grouping all the descriptors into a set of clusters using k -means. Bag-of-features is formed by calculating the number of the descriptor in each cluster. This resulting fully connected feature vector is merged with the last fully connected layer of CNN.

3.2.1 PCA to reduce dimension of Handcrafted feature SIFT

However, the resulting feature vector extracted from more number sequential frame has increased the dimensionality of motion features. Again, we need to concatenate the SIFT feature vector with the last layer or classification layer of CNN, which produce a giant vector for classification. The high dimensionality of these combined feature vector will become very complex for classification. To address this problem, the subspace method proposed in Nguyen et al. (2017) is used to reduce the dimension of handcrafted SIFT features before the concatenation of features vectors. We have used PCA (principal component analysis) to select less number of components for extracted SIFT features before concatenating with classification layer of CNN as shown in Fig. 1. By using PCA, we can select less number of principal components having maximum variation from original components. The resulting principal components can have sufficient power to represent the original features with a smaller dimension.

3.3 Deep feature extraction using CNN

2D CNN computes features in one dimension(spatial dimension), and it can extract very useful features from single frame still image. If we talk about video analysis, we need to capture the motion information between consecutive frames; therefore, we have to compute features from both the dimension(spatial and temporal). 3D CNN is introduced by applying a 3D kernel to the contiguous input video frames (Ji et al. 2013). By this convolution, the obtained feature maps are connected to multiple frames of the input video, which

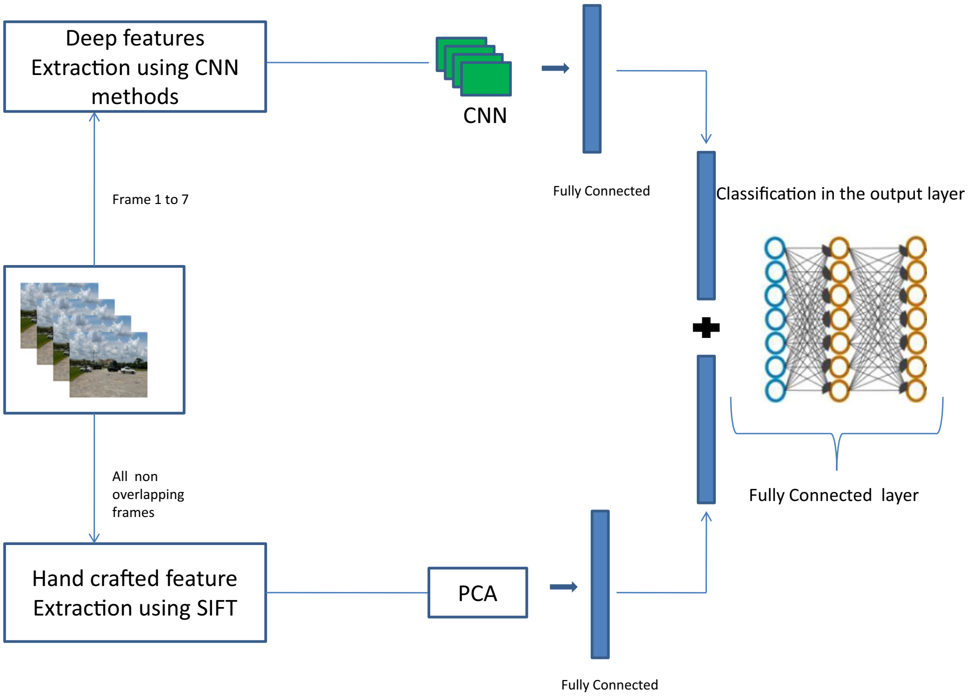

Deep features Extraction using CNN methods
CNN
Fully Connected
Frame 1 to 7
Classification in the output layer
All non overlapping frames
Fully Connected layer
Hand crafted feature Extraction using SIFT
PCA
Fully Connected

Fig. 1 Overview of the proposed framework

may capture the motion information effectively. Mathematical representation of 3D CNN is shown in Eq. 1.

$$v_{ij}^{xy} = \tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{p_1-1}\sum_{q=0}^{q_1-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right) \tag{1}$$

where v_{ij}^{xy} represent value at position (x, y) in ith layer of jth feature map. $tanh$ is hyperbolic tangent function, b_{ij} is bias, m is indexes, w_{ijm}^{pq} is the value at position (p, q) of the kernel and P_i and Q_i are height and width of the kernel.

3.4 3D CNN on video frames

In order to capture deep spatial and temporal motion features from action video, we adopt the state-of-the-art 3D CNN architecture proposed by Ji et al. (2013) and have done few modifications to combined other handcrafted features in the classification layer. In our architecture shown in Fig. 3, we have taken seven contagious frames of size 60 × 40 as input to our model. Instead of random initializing input frames, we have taken features like gradient X and Y and optical flow X and Y of all the seven frames to produce several channels of information from the input video. The resulting channels of information act as the feature map for the next layer. These hardwired features are employed before the convolution operation to encode human knowledge on features. This handcrafted information has shown effective performance than providing direct raw input to the network. Gradient X and Y and optical flow in both X and Y directions are explained below.

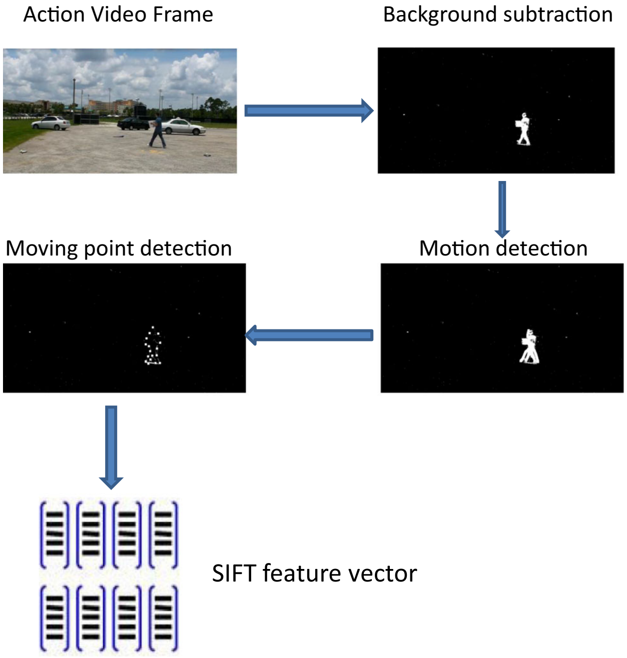


Fig. 2 Description of SIFT feature extraction

3.4.1 Gradient of an image (sobel filter)

The gradient of an image is used to extract meaningful information from an image. X and Y gradient images are generated from the original image by convolving with a filter. It measures the change in intensity of any pixel of an image with direction. First, seven frames of size 60 × 40 are taken from

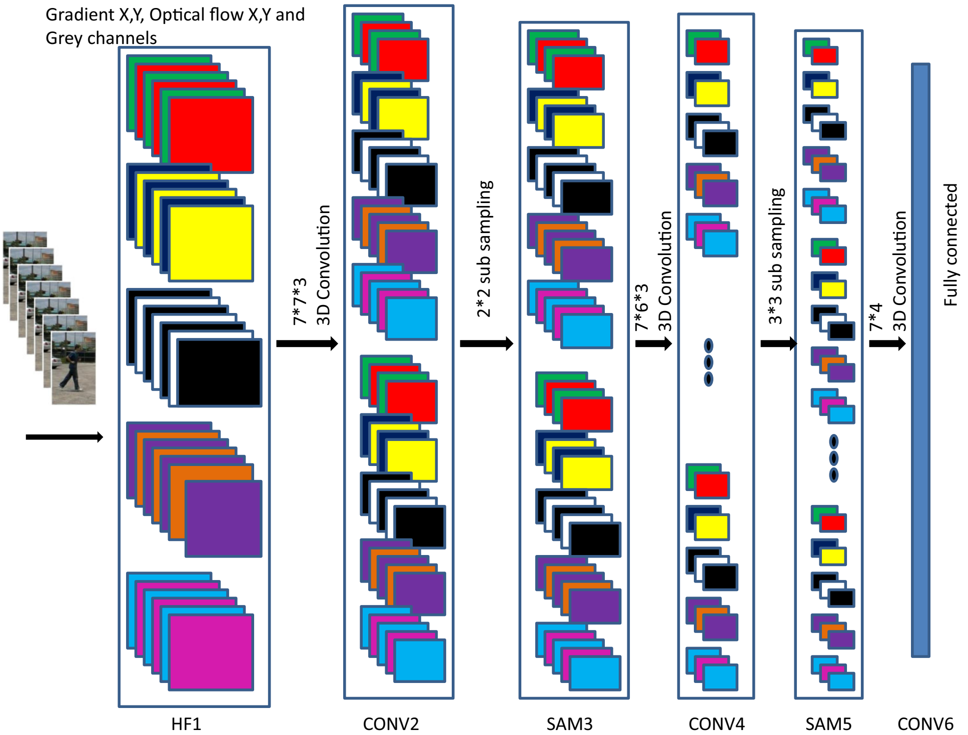


Fig. 3 3D CNN architecture for deep feature extraction

the input video, and gradient in X and Y direction is computed from the input frames. Then instead of raw input, these gradients of all input frames are given to the network for convolution operation. Gradients representation in X and Y direction of the input image is shown in Fig. 4.

3.4.2 Optical flow

Optical flow between two consecutive video frames is the motion of any pixel location, caused by movement of any objects in a scene. For example in first frame the intensity of pixel is $I(x, y, t)$ at time t, if we move its pixel by $(\mathrm{d}x, \mathrm{d}y)$ at time dt, then we get new pixel intensity $I(x + \mathrm{d}x, y + \mathrm{d}y, t + \mathrm{d}t)$. In optical flow calculation, we have to assume image pixel intensity are constant in all frames. With this assumption, we can form a equation $I(x, y, t) = I(x + \mathrm{d}x, y + \mathrm{d}y, t + \mathrm{d}t)$. Then, we take the Taylor Series Approximation of the right-hand side to get the below equation.

$$I(x + \delta x, y + \delta y, t + \delta t) = \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t} \tag{2}$$

Now, divide dt to get the final optical flow equation below.

$$= \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} \tag{3}$$

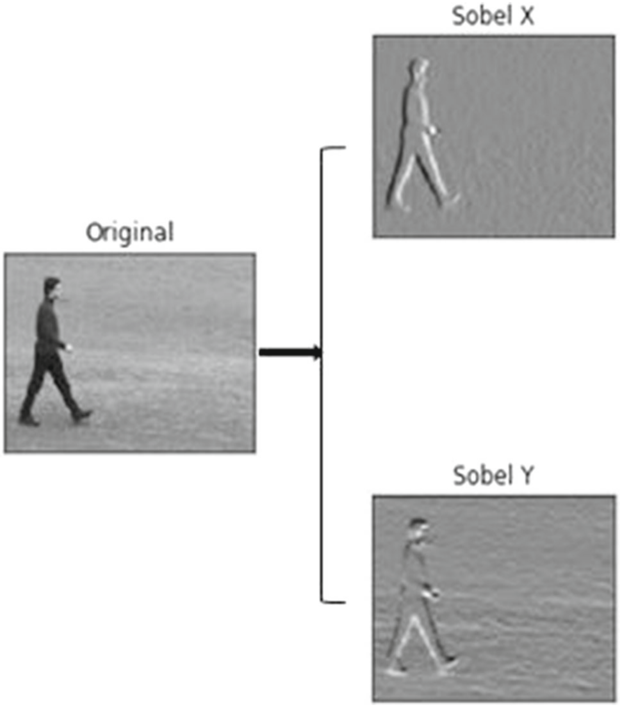


Fig. 4 Gradient representation in X and Y direction

where $u = \dfrac{\mathrm{d}x}{\mathrm{d}t}$ and $v = \dfrac{\mathrm{d}y}{\mathrm{d}t}$. Here, u and v are the movements over time t. u and v can be solve using Lucas–Kanade method. This optical flow is very much effective in motion sequence analysis of human action in video data. We have

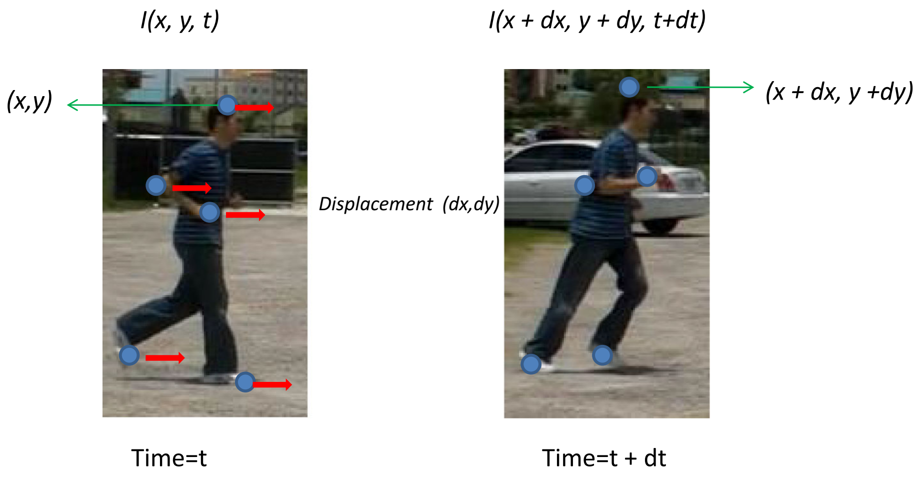


Fig. 5 Optical flow between two consecutive frames in action video

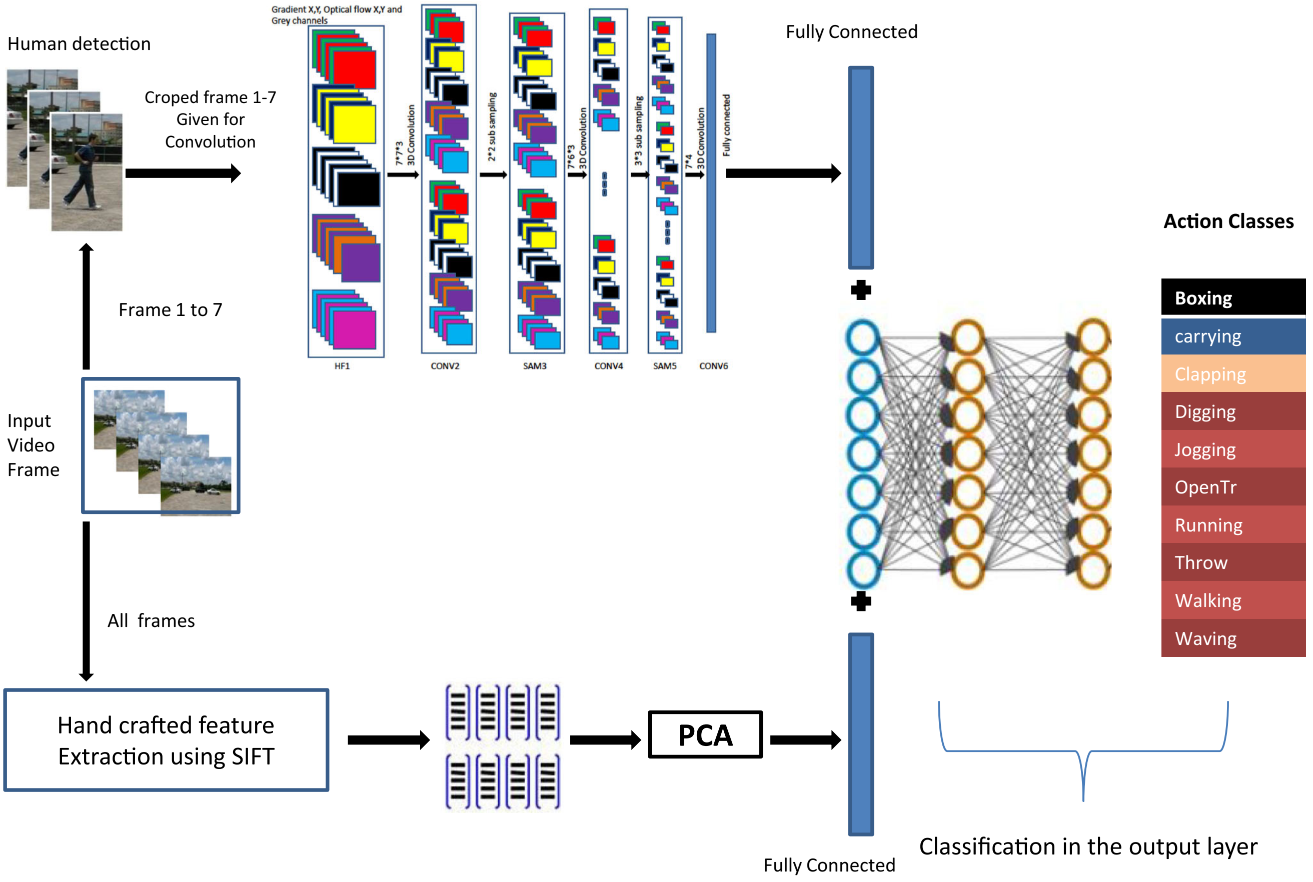


Fig. 6 Proposed SIFT-CNN architecture

calculated the gradient and optical flow of seven consecutive frames and generated multiple information channels before the 3D convolution process. The optical flow calculation of frames is shown in Fig. 5.

3.5 Proposed convolution operation

Several CNN architectures can be invented from the 3D convolution technique. As described above, we have extracted gradient X, Y, gray channel, and optical flow for all the frames from the input video in the first layer. After being processed through these feature extraction phases, it generates multiple information channels in the form of feature maps in the second layer. The feature maps contain gradient along both horizontal and vertical direction and optical flow field in both direction X and Y and gray pixel value for all input frames. Next, we apply convolutions operation in both spatial

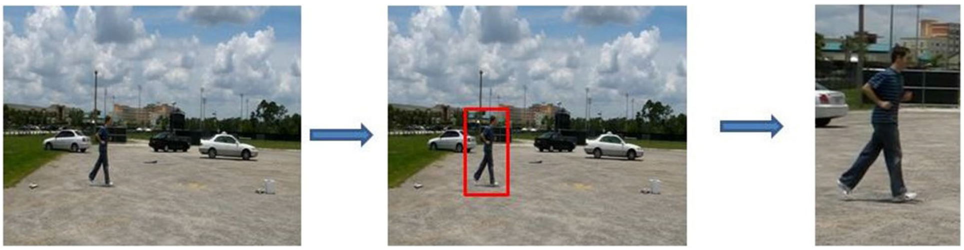


Fig. 7 Human detection with bounding box from original frames in UCF dataset

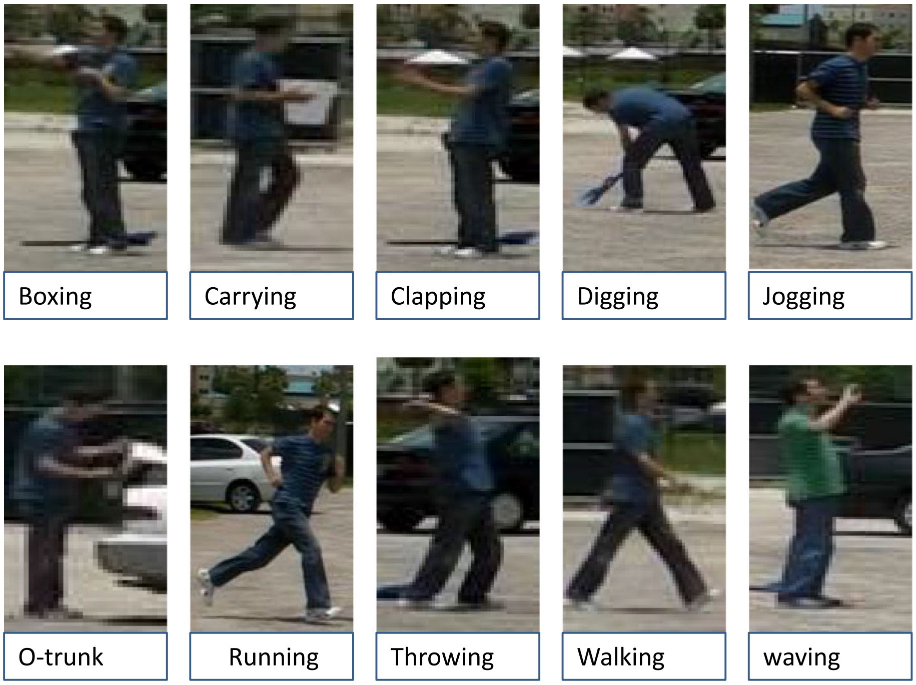


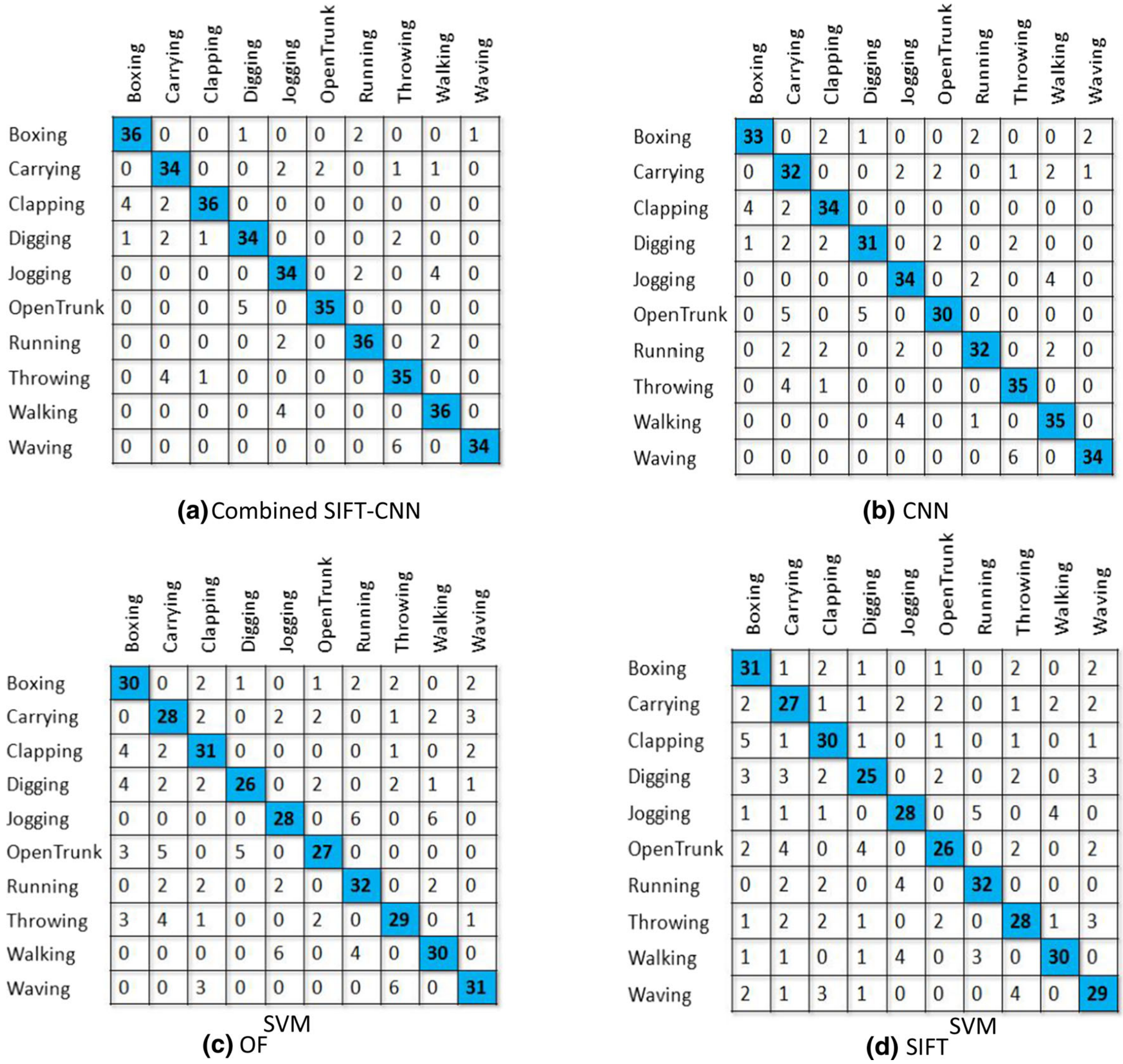
Fig. 8 Sample human action from UCF dataset

and temporal domain on all these extracted features channels with a kernel size of 7 × 7 × 3, 7 × 7 in spatial and 3 in the temporal dimension, respectively. Two rounds of convolution operation are applied on every channel separately to get the maximum number of feature maps. Now, 2 × 2 subsampling is applied on obtained two sets of feature maps from the previous convolution. Subsampling layers are used to reduce only the spatial resolution by keeping the same number of feature maps. Again 3D convolution is applied on these two sets of sampled feature maps with a kernel size of 7 × 6 × 3 on all resulting channels. Here, we have applied three different kernel combinations on each of the two sets of features map and obtained six sets of feature maps. Again 3 × 3 subsampling is done for all these resulting feature maps to reduce the spatial dimension. Now, the size of the temporal dimension is reduced, so the next convolution is applied only in spatial dimension with a kernel size of 7 × 4. As a result, final feature maps of size 1 × 1 were obtained as a fully connected layer.

Python and TensorFlow are used to implement the proposed CNN on UCF and KTH datasets. The detailed implementation is based on 3D CNN proposed in Ji et al. (2013). In convolution layer, 7 × 7 × 3 filters are used to extract subregion with ReLU activation function. Subsampling is done in the pooling layer. After several rounds of convolution and pooling, we obtained 128 dimensions fully connected feature vectors from 7 input frames. All parameters used in this model are taken randomly, and training is done by using the backpropagation algorithm proposed in LeCun et al. (1998).

3.6 SIFT and CNN aggregator

It is already admitted that the human visual system is a multiscale process (Donoho and Huo 2002). Therefore, it will be advisable to integrate multiple levels of motion features for the robust classification of human action. The combination of invariant and more delicate features produces a better representation of motion features. As shown in Fig. 6, we have extracted two feature vectors (Using SIFT and CNN) that describe the input video's spatial and temporal information. In the above CNN architecture, we have used only seven input frames for convolution operation, which is insufficient to encode high-level temporal motion information from complete action video. Again, if we increase the input frame, handling all the trainable parameters of the network will be



	Boxing	Carrying	Clapping	Digging	Jogging	OpenTrunk	Running	Throwing	Walking	Waving
Boxing	36	0	0	1	0	0	2	0	0	1
Carrying	0	34	0	0	2	2	0	1	1	0
Clapping	4	2	36	0	0	0	0	0	0	0
Digging	1	2	1	34	0	0	0	2	0	0
Jogging	0	0	0	0	34	0	2	0	4	0
OpenTrunk	0	0	0	5	0	35	0	0	0	0
Running	0	0	0	0	2	0	36	0	2	0
Throwing	0	4	1	0	0	0	0	35	0	0
Walking	0	0	0	0	4	0	0	0	36	0
Waving	0	0	0	0	0	0	0	6	0	34

(a) Combined SIFT-CNN

	Boxing	Carrying	Clapping	Digging	Jogging	OpenTrunk	Running	Throwing	Walking	Waving
Boxing	33	0	2	1	0	0	2	0	0	2
Carrying	0	32	0	0	2	2	0	1	2	1
Clapping	4	2	34	0	0	0	0	0	0	0
Digging	1	2	2	31	0	2	0	2	0	0
Jogging	0	0	0	0	34	0	2	0	4	0
OpenTrunk	0	5	0	5	0	30	0	0	0	0
Running	0	2	2	0	2	0	32	0	2	0
Throwing	0	4	1	0	0	0	0	35	0	0
Walking	0	0	0	0	4	0	1	0	35	0
Waving	0	0	0	0	0	0	0	6	0	34

(b) CNN

	Boxing	Carrying	Clapping	Digging	Jogging	OpenTrunk	Running	Throwing	Walking	Waving
Boxing	30	0	2	1	0	1	2	2	0	2
Carrying	0	28	2	0	2	2	0	1	2	3
Clapping	4	2	31	0	0	0	0	1	0	2
Digging	4	2	2	26	0	2	0	2	1	1
Jogging	0	0	0	0	28	0	6	0	6	0
OpenTrunk	3	5	0	5	0	27	0	0	0	0
Running	0	2	2	0	2	0	32	0	2	0
Throwing	3	4	1	0	0	2	0	29	0	1
Walking	0	0	0	0	6	0	4	0	30	0
Waving	0	0	3	0	0	0	0	6	0	31

SVM
(c) OF

	Boxing	Carrying	Clapping	Digging	Jogging	OpenTrunk	Running	Throwing	Walking	Waving
Boxing	31	1	2	1	0	1	0	2	0	2
Carrying	2	27	1	1	2	2	0	1	2	2
Clapping	5	1	30	1	0	1	0	1	0	1
Digging	3	3	2	25	0	2	0	2	0	3
Jogging	1	1	1	0	28	0	5	0	4	0
OpenTrunk	2	4	0	4	0	26	0	2	0	2
Running	0	2	2	0	4	0	32	0	0	0
Throwing	1	2	2	1	0	2	0	28	1	3
Walking	1	1	0	1	4	0	3	0	30	0
Waving	2	1	3	1	0	0	0	4	0	29

SVM
(d) SIFT

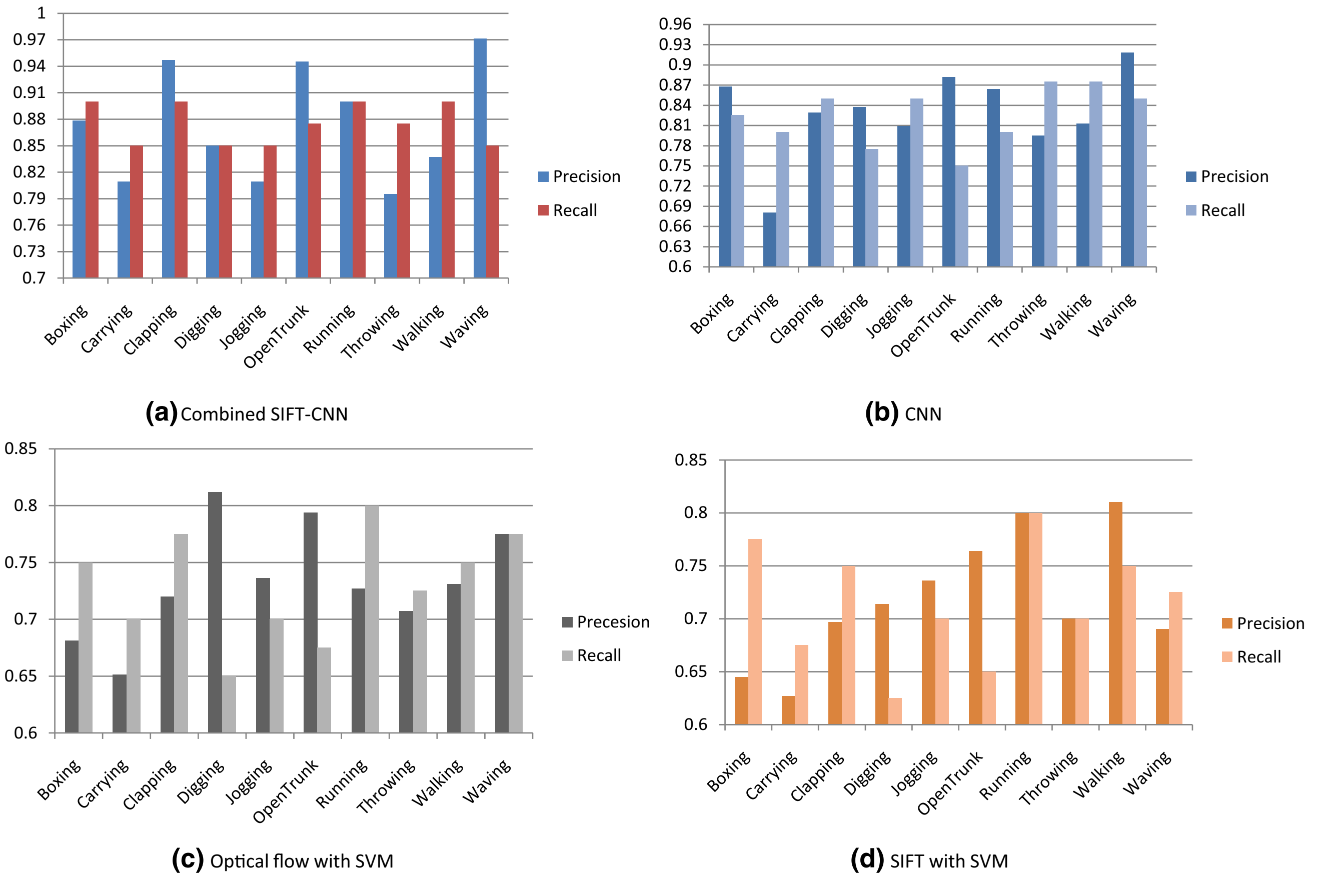
Fig. 9 Confusion matrix for all 4 models

very complex. To address this problem, we have used key points-based descriptor SIFT in parallel with CNN to capture the motion information of all the frames present in the action video. In this approach, SIFT features is considered as spatially distinctive interest points and the displacement of these distinctive points is computed to find motion constrain around it. In the above CNN architecture, there is a global average pooling before the fully connected layer. We pass the extracted SIFT into the dense layer and concatenate CNN's fully connected layer with SIFT feature vector. After concatenation of two feature vectors, one more fully connected layer followed by a dropout of 0.4 and one more dense layer with sigmoid activation function is added for final classification. To avoid overfitting of the network, All pooling and dense layers are followed by a dropout layer. All training was performed by using Caffe (Vedaldi et al. 2014) framework with an additional feature vector SIFT at the classification layer. Dropouts of 0.4 and 0.6 were tested, and we found 0.6 gave a better result in our proposed combined framework. Momentum and Weight decay were set to 0.8 and 0.0005 to reduce the overfitting of the network. We adopt the approach in Simonyan and Zisserman (2014b) for learning rate and testing purposes.

4 Experiments

4.1 Action recognition on UCF dataset

We have used UCF and KTH datasets to evaluate the proposed combined CNN model for human action recognition. UCF Human action dataset consists of 10 actions performed by 12 actors recorded from 3 different angles. We have taken the aerial view recording for our experiment. The 10 actions are Boxing, Clapping, Digging, Jogging, Carrying, Open-Close Trunk, Running, Throwing, Walking, and Waving,



(a) Combined SIFT-CNN

(b) CNN

(c) Optical flow with SVM

(d) SIFT with SVM

Fig. 10 Precision Recall bar-chart for all 4 models

shown in Fig. 4. All aerial view recording videos contain human action along with multiple objects with different backgrounds. Therefore, we first used a human detector to detect and prepare bounding boxes from 960 × 540 video frames and then reduced to 60 × 40 shown in Fig. 7 for convolution operation. Human detection with bounding box from original frames is explained in Yang et al. (2009) (8.

As we have taken 7 input frames for the convolution process, so the above detection and tracking process is applied for all the 7 frames to get a cube having only human action without background. We have extracted humans with bounding boxes with a step size of 2 from the original frames. For example, for frames number 0, bounding boxes are extracted for frames numbers −2, −4, −6, 2, 4, 6. For all these frames, bounding boxes are extracted at the same position. Finally, the action video of size 960 × 540 is converted to 60 × 40 patch cube, which is the input to our model in one stream. Along with this convolution process, we have computed SIFT for the same input 960 × 540 in parallel with CNN. SIFT descriptor is extracted around every detected point in both spatial and temporal dimension for all the frames present in the input video clip, as explained in Sect. 3.2. Finally, PCA is applied to these features vectors to reduce the dimension and concatenated with the last layer of the proposed CNN.

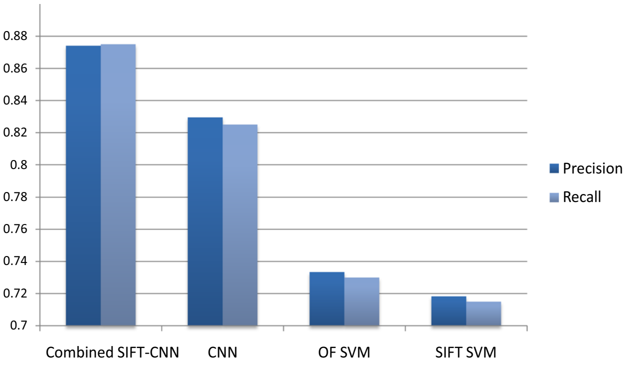


Fig. 11 Average Precision recall

To evaluate the performance of the combined CNN framework shown in Fig. 6, we have reported the result of standard 3D CNN along with OpticalFlow(OF) and SVM, SIFT, and SVM on the UCF dataset. These are represented as OF^{SVM} and $SIFT^{SVM}$. In OF^{SVM}, the optical flow descriptor is generated, and K mean clustering is used with the different number of clusters. These clusters are used for vector quan-

Table 1 Performance(Precision and Recall) for all 4 methods

		Boxing	Carrying	Clapping	Digging	Jogging	OpenTrunk	Running	Throwing	Walking	Waving	Avg
Combined SIFT-CNN	Precision	0.878	0.809	0.947	0.85	0.809	0.945	0.9	0.795	0.837	0.971	0.8741
	Recall	0.9	0.85	0.9	0.85	0.85	0.875	0.9	0.875	0.9	0.85	0.875
CNN	Precision	0.868	0.68	0.829	0.837	0.809	0.882	0.864	0.795	0.813	0.918	0.8295
	Recall	0.825	0.8	0.85	0.775	0.85	0.75	0.8	0.875	0.875	0.85	0.825
Optical flow with SVM	Precision	0.681	0.651	0.72	0.812	0.736	0.794	0.727	0.707	0.731	0.775	0.7334
	Recall	0.75	0.7	0.775	0.65	0.7	0.675	0.8	0.725	0.75	0.775	0.73
SIFT with SVM	Precision	0.645	0.627	0.697	0.714	0.736	0.764	0.8	0.7	0.81	0.69	0.7183
	Recall	0.775	0.675	0.75	0.625	0.7	0.65	0.8	0.7	0.75	0.725	0.715

Table 2 Comparison to state-of-the-art method on UCF dataset

Models	Accuracy
Two Stream with extra dataSimonyan and Zisserman (2014b)	86.9
Two stream, regularized fusionWu et al. (2015)	88.4
Two stream with extra data, SVM fusionSimonyan and Zisserman (2014b)	88.0
CNN, optical flow, LSTMYue-Hei Ng et al. (2015)	88.6
CNN, IDT, FVZha et al. (2015)	89.6
IDT, FV, temporal scale invarianceLan et al. (2015)	89.1
Ours	89.5

tization to assign each optical flow descriptor to its nearest codeword. Then, bag of word(BOW) vector is constructed for each video. A BOW vector is like a histogram that counts the frequency of optical flow descriptors that appears in the video. Then, we have used Linear SVM to train these vectors, which contain different human actions. Similarly, in $SIFT^{SVM}$, SIFT features extraction, K-means clustering to build codebook, Building of Bag-of-Words vector, Training with SVM are done step by step for every video of the training set.

We have used 40 random samples of UCF dataset and report the performance of these 4 models with the help of the confusion matrix shown in Fig. 9. We can clearly observe from confusion matrices that most of the actions are correctly classified in Fig. 9a compared to Fig 9b, c, and d. We can also observe, our combined CNN outperforms the normal 3D CNN, OF^{SVM} and $SIFT^{SVM}$. In few cases, combined CNN and general 3D CNN are showing the same results; otherwise, our proposed combined CNN is showing better performance for all actions. Precision and recall are computed for all these models with 10 different actions. The comparative bar-chart is shown in Fig. 10 for all 4 models. We can clearly observe from the bar-chart and Table 1 that combined CNN is showing better precision and recall compared to others. Average precision and recall of 10 human actions also plotted for all models is shown in Fig. 11.

We have tested the time complexity of handcrafted SIFT feature extraction from the input video and observed that point detection and description time is much faster. The time-space for histogram formation is also very less. The key point detection algorithm generates fewer feature points, and it detects only moving points from the 2D frames. This helps extract meaningful information from the maximum number of frames present in the action video with reduced computational complexity. Table 2 shows state-of-the-art method

Table 3 The Recognition accuracy in percentage for KTH dataset

Methods	Boxing	Hand clapping	Hand waving	Jogging	Running	Walking	AVG
Combined SIFT-CNN	91	92	95	84	82	96	90
CNN	87	90	93	86	74	84	85.6
OF SVM	85	84	86	74	90	72	81.8
SIFT SVM	86	79	84	71	84	70	79

for human action recognition on UCF dataset(Simonyan and Zisserman 2014b; Wu et al. 2015; Yue-Hei Ng et al. 2015; Zha et al. 2015; Lan et al. 2015). We achieved comparatively better result by considering longer temporal information with less computational complexity and time.

4.2 Action recognition On KTH dataset

We also evaluate all these 4 models with KTH human action data, consisting of 6 actions performed by 25 different subjects. We have used a cube of total 9 frames as input after foreground extraction. 80 × 60 resolution is used in these experiments for fewer memory requirements. We have used the same 3D CNN architecture as in Fig. 3. As the resolution of images is 80 × 60 and 9-frame are used, so convolution layer kernel sizes are $9 \times 7, 7 \times 7, 6 \times 4$ and subsampling layer kernel sizes are 3 × 3. After multiple layers of convolution and subsampling, finally, we get a flat, fully connected feature vector. The last layer contains 6 units. The same setting is used for combined CNN shown in Fig. 6 with additional features SIFT parallel with CNN. 16 random subjects are selected for training and 9 used for testing. The same training and testing split are used for all 4 methods and observed that our combined CNN achieves excellent accuracy compared to the remaining three shown in Table 3.

5 Conclusion

In this paper, we proposed a combined SIFT and CNN model for human action recognition. This model extracts features from both spatial and temporal dimension like normal 3D convolution. This model also extracts additional handcrafted features from the entire input video clip(All frames). Finally, the last fully connected layer of the CNN is concatenated with the additional handcrafted feature vectors SIFT. These combinations of feature vectors preserved both spatial and longer temporal information from the entire video clip without increasing CNN's input window and boosted the model performance. We evaluated the combined CNN on UCF and KTH datasets. We observed that the combination of handcrafted features with 3D CNN outperforms both UCF and KTH datasets with a fixed number of training and testing splits. We compared our model with normal 3D CNN, OF with SVM, and SIFT with SVM with a fixed number of a random sample from both datasets. Comparison to state-of-the-art method on UCF dataset also presented. However, we proposed to combine deep and handcrafted features in the classification layer of the SIFT-CNN framework. There is no score level fusion, and weight values are considered for combining two different types of features. This could be the drawback of our proposed model because the score level fusion with fixed weight may significantly increase the accuracy of the network. Nowadays, depth sensor RGB-D data received good attention in human action recognition, and commercial depth sensors are easily available. We believe that our proposed framework improves action recognition performance by combining handcrafted depth RGB data and static images. We will explore the same combination strategy with RGB-D data in our future work.

Author Contributions All authors have equally contributed, and all authors have read and agreed to the published version of the manuscript.

Funding This research has no funding by any organization or individual.

Data Availability Data sharing not applicable.

Declaration

Conflict of interest The authors declare no conflict of interest.

References

Algamdi AM, Sanchez V, Li CT (2019) Learning temporal information from spatial information using capsnets for human action recognition. In: ICASSP 2019—2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 3867–3871. IEEE

Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis & Machine Intelligence 3:257–267

Chakraborty B, Holte MB, Moeslund TB, Gonzàlez J (2012) Selective spatio-temporal interest points. Computer Vision and Image Understanding 116(3):396–410

Choutas V, Weinzaepfel P, Revaud J, Schmid C (2018) Potion: Pose motion representation for action recognition. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Dawn DD, Shaikh SH (2016) A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector. The Visual Computer 32(3):289–306

Deng L, Wang Y, Liu B, Liu W, Qi Y (2018) Biological modeling of human visual system for object recognition using glop filters and sparse coding on multi-manifolds. Machine Vision and Applications 29(6):965–977

Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrell T (2017) Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell 39(4):677–691

Donoho DL, Huo X (2002) Beamlets and multiscale image analysis. In: Multiscale and multiresolution methods, pp. 149–196. Springer

Feichtenhofer C, Pinz A, Wildes RP (2017) Spatiotemporal multiplier networks for video action recognition. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR). IEEE

Gaidon A, Harchaoui Z, Schmid C (2014) Activity representation with motion hierarchies. International journal of computer vision 107(3):219–238

Garain J, Mishra SR, Kumar RK, Kisku DR, Sanyal G (2019) Bezier cohort fusion in doubling states for human identity recognition with multifaceted constrained faces. Arabian Journal for Science and Engineering 44(4):3271–3287

Jhuang H, Serre T, Wolf L, Poggio T (2007) A biologically inspired system for action recognition. In: International conference on computer vision (ICCV)

Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231

Junejo IN, Dexter E, Laptev I, Pérez P (2011) View-independent action recognition from temporal self-similarities. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(1):172–185

Kawai Y, Takahashi M, Fujii M, Naemura M, Satoh S (2010) Nhk strl at trecvid 2010: semantic indexing and surveillance event detection. In: TRECVID

Khaire P, Kumar P, Imran J (2018) Combining cnn streams of rgb-d and skeletal data for human activity recognition. Pattern Recognition Letters 115:107–116

Klaser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3D-gradients

Kong Y, Tao Z, Fu Y (2017) Deep sequential context networks for action prediction. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE

Lan Z, Lin M, Li X, Hauptmann AG, Raj B (2015) Beyond gaussian pyramid: multi-skip feature stacking for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 204–212

Laptev L (2003) Space-time interest points. In: Proceedings ninth IEEE international conference on computer vision. IEEE

Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2–3):107–123

Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR 2011. IEEE

LeCun Y, Bottou L, Bengio Y, Haffner P et al (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324

Li H, Li H, Wei Y, Tang Y, Wang Q (2014) Sparse-based neural response for image classification. Neurocomputing 144:198–207

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2):91–110

Ma M, Marturi N, Li Y, Leonardis A, Stolkin R (2018) Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos. Pattern Recognition 76:506–521

Mishra SR, Krishna KD, Sanyal G, Sarkar A et al (2020) A feature weighting technique on svm for human action recognition. Journal of Scientific and Industrial Research (JSIR) 79(7):626–630

Mishra SR, Mishra TK, Sanyal G, Sarkar A, Satapathy SC (2020) Real time human action recognition using triggered frame extraction and a typical cnn heuristic. Pattern Recognition Letters 135:329–336

Nazir S, Yousaf MH, Velastin SA (2018) Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition. Computers & Electrical Engineering 72:660–669

Núñez JC, Cabido R, Pantrigo JJ, Montemayor AS, Vëlez JF (2018) Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition 76:80–94

Nguyen, D., Kim, K., Hong, H., Koo, J., Kim, M., Park, K.: Gender recognition from human-body images using visible-light and thermal camera videos based on a convolutional neural network for image feature extraction. Sensors 17(3), 637 (2017)

Nguyen TV, Song Z, Yan S (2014) Stap: Spatial-temporal attention-aware pooling for action recognition. IEEE Transactions on Circuits and Systems for Video Technology 25(1):77–86

Ning F, Delhomme D, LeCun Y, Piano F, Bottou L, Barbano P (2005) Toward automatic phenotyping of developing embryos from videos. IEEE Trans Image Process 14(9):1360–1371

Patel CI, Garg S, Zaveri T, Banerjee A, Patel R (2018) Human action recognition using fusion of features for unconstrained video sequences. Computers & Electrical Engineering 70:284–301

Peng X, Wang L, Wang X, Qiao Y (2016) Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding 150:109–125

Peng X, Zou C, Qiao Y, Peng Q (2014) Action recognition with stacked fisher vectors. In: European conference on computer vision. pp. 581–595. Springer

Ramezani M, Yaghmaee F (2016) A review on human action analysis in videos for retrieval applications. Artificial Intelligence Review 46(4):485–514

Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th international conference on neural information processing systems, Vol 1, pp. 568–576. MIT Press

Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576

Tomasi C, Detection TK (1991) Tracking of point features. Tech. rep., Tech. Rep. CMU-CS-91-132, Carnegie Mellon University

Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE international conference on computer vision (ICCV). IEEE

Tran D, Ray J, Shou Z, Chang SF, Paluri M (2017) Convnet architecture search for spatiotemporal feature learning. arXiv preprint arXiv:1708.05038

Tu Z, Xie W, Qin Q, Poppe R, Veltkamp RC, Li B, Yuan J (2018) Multi-stream cnn: Learning representations based on human-related regions for action recognition. Pattern Recognition 79:32–43

Vedaldi A, Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Darrell T (2014) Convolutional architecture for fast feature embedding. Cornell University, arXiv:1408.5093 v12014

Wang H, Oneata D, Verbeek J, Schmid C (2016) A robust and efficient video representation for action recognition. International Journal of Computer Vision 119(3):219–238

Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision. pp. 3551–3558

Wang L, Koniusz P, Huynh DQ (2019) Hallucinating bag-of-words and fisher vector IDT terms for CNN-based action recognition. arXiv preprint arXiv:1906.05910

Wang Y, Mori G (2011) Hidden part models for human action recognition: Probabilistic versus max margin. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(7):1310–1323

Wu Z, Wang X, Jiang YG, Ye H, Xue X (2015) Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In: Proceedings of the 23rd ACM international conference on multimedia. pp. 461–470. ACM

Yang M, Lv F, Xu W, Gong Y (2009) Detection driven adaptive multi-cue integration for multiple human tracking. In: 2009 IEEE 12th international conference on computer vision. IEEE

Yu K, Xu W, Gong Y (2009) Deep learning with kernel regularization for visual recognition. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Advances in neural information processing systems, vol 21. Curran Associates Inc, New York, pp 1889–1896

Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4694–4702

Zha S, Luisier F, Andrews W, Srivastava N, Salakhutdinov R (2015) Exploiting image-trained CNN architectures for unconstrained video classification. arXiv preprint arXiv:1503.04144

Zhu W , Lan C, Xing J, Zeng W, Li Y, Shen L, Xie X (2016) Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: Thirtieth AAAI conference on artificial intelligence

Zhang Z, Hu Y, Chan S, Chia LT (2008) Motion context: A new representation for human action recognition. In: European conference on computer vision. pp. 817–829. Springer

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