Robust video content analysis schemes for human action recognition

Cherry A. Aly, Fazly S. Abas and Goh H. Ann
Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia

Abstract

Introduction: Action recognition is a challenging time series classification task that has received much attention in the recent past due to its importance in critical applications, such as surveillance, visual behavior study, topic discovery, security, and content retrieval.

Objectives: The main objective of the research is to develop a robust and high-performance human action recognition techniques. A combination of local and holistic feature extraction methods used through analyzing the most effective features to extract to reach the objective, followed by using simple and high-performance machine learning algorithms.

Methods: This paper presents three robust action recognition techniques based on a series of image analysis methods to detect activities in different scenes. The general scheme architecture consists of shot boundary detection, shot frame rate re-sampling, and compact feature vector extraction. This process is achieved by emphasizing variations and extracting strong patterns in feature vectors before classification.

Results: The proposed schemes are tested on datasets with cluttered backgrounds, low- or high-resolution videos, different viewpoints, and different camera motion conditions, namely, the Hollywood-2, KTH, UCF11 (YouTube actions), and Weizmann datasets. The proposed schemes resulted in highly accurate video analysis results compared to those of other works based on four widely used datasets. The First, Second, and Third Schemes provides recognition accuracies of 57.8%, 73.6%, and 52.0% on Hollywood2, 94.5%, 97.0%, and 59.3% on KTH, 94.5%, 95.6%, and 94.2% on UCF11, and 98.9%, 97.8% and 100% on Weizmann.

Conclusion: Each of the proposed schemes provides high recognition accuracy compared to other state-of-art methods. Especially, the Second Scheme as it gives excellent comparable results to other benchmarked approaches.

Keywords
Action recognition, video analysis, binary robust invariant scalable keypoints, bag of words, histogram of oriented gradient, histogram of optical flow, space-time interest points

Corresponding author:
Cherry Aly, Faculty of Engineering and Technology, Multimedia University, Jalan Ayer Keroh Lama, Melaka 75450, Malaysia.
Email: cherry_ali_eng@hotmail.com

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
Introduction

Modern-day smartphones and multimedia devices have made it possible for any individual to capture, share, and receive digital videos over the internet. Statistics show that on a single online video platform, that is, YouTube, 300 h of video content is uploaded every minute. Besides, there are thousands of surveillance and security cameras installed in cities that produce videos 24 h a day. With such a large video content being produced every second, it becomes a very hectic task to evaluate and analyze these videos for security and censorship purposes. This wide expansion of online video content has also created a significant demand for intelligent video analysis and parsing methods. Here, the action recognition method plays a vital role in analyzing video content.

Action recognition is a challenging task as it involves analyzing and predicting specific human movements in the video. Action recognition has a large number of very important applications including location estimation, goal recognition, human-computer interactions, and sociology. Generally, action recognition techniques are based on the following two steps: (1) feature extraction and (2) action classification.

In this paper, robust action recognition techniques benchmarked against existing ones are presented. Improvements are demonstrated through highly accurate results achieved on the datasets Hollywood2, KTH, UCF11, and Weizmann. The present study combines the two major groups of feature extraction and proposed action recognition techniques; these groups are local feature-based (LFB) and holistic feature-based (HFB). The LFB includes cuboid detector, 3D Harris corner detector, and 3D scale-invariant feature transform (3D SIFT), which tends to increase robustness against a complex background and occlusion in realistic actions.

The HFB includes a histogram of 3D-oriented gradient (3D HOG), histogram of optical flow (HOF), motion history image, 3D convolutional neural networks (3D CNN), sparse image coding, regions of interests (ROI), etc., which use the visual information to represent actions from the whole sequence of frames. The holistic representation is considered sensitive to photometric distortions and geometric shifting.

The proposed schemes progress using shot boundary detection and/or frame rate resampling. A final compact feature vector of resampled frames representing each shot and concatenated alignments of the videos are incorporated at different classification models such as Support Vector Machines (SVM), Bagged Trees, and K-Nearest Neighbor.

We briefly emphasis our main research contributions as follows:

- Due to increasing of volume of data, we tried to minimize the input video size by detecting the shot boundary using a simple algorithm to focus on the actions in each shot and as a minimizing pre-processing for the next step.
• As a continuity of eliminating the video data size process, we applied a sub-sampling technique on each detected shot by selecting some frames to represent each shot.

• Due to the enormous variety of data and its challenges such as camera motion, pose, crowded background, and different lighting conditions that all occur in the used four datasets we tried to apply the most effective scenario of well-reconstructed features from our perspective to achieve our goal of high accuracy recognition.

• From the feature representation perspective, our contribution lies in finding the most compact feature vector keeping the most powerful patterns representing each video in a low-dimensional output vector.

• We are using more than one classification model to compromise the best one classified action results with our propose three schemes.

The remainder of this paper is structured as follows: Section II presents a survey of the relevant literature. Section III describes the details of proposed schemes for action recognition. Section IV focuses on experimental results and their comparison with recently published research articles on action recognition. Section V is a conclusion of the present research.

Relevant literature

The complexity of video analysis and human action recognition has received significant attention in recent decades. Initially, video recognition was not possible due to the low processing power available; however, with the debauched evolution of modern-day high-power computers and smart algorithms, it is becoming easy for the software to analyze and understand actions in a video. However, action recognition is still considered as an open challenging task due to the immense variation in video content, lighting conditions, camera positions, and the like.

Several studies have demonstrated significant progress in this field, and this paper considers the latest methods presented for action recognition. In one such study, Sekma et al.\textsuperscript{13} proposed a structured Fisher vector encoding technique using geometric relationships among trajectories and SVM for classification. Similarly, Liu et al.\textsuperscript{14} proposed a genetic programming evolutionary method that evolved from the motion feature descriptors (3D Gaussian, 3D Laplacian, 3D Gabor, and 3D Wavelet) and SVM for classification. Kardaris et al.\textsuperscript{15} compose a dominant visual word sequence to represent actions and apply local sub-sequence alignment to assess the sequence’s similarity using a Smith-Waterman local sequence alignment algorithm and then using SVM for classification. Similarly, Xu et al.\textsuperscript{16} proposed a “dictionary learning architecture” of two streams, which consists of an interest patch descriptor and detector as well as support vector machine for classification.

Studies have shown that a convolutional neural network (CNN) is more successful at parsing videos for action recognition. In one such study, Shi et al.\textsuperscript{17} used
the sequential deep trajectory descriptor (SDTD) for action recognition into three-dimensional planes and employed CNN for effective short- and long-term motion representation. Zhang et al.\textsuperscript{18} utilized an image-trained two-layer convolutional neural network (CNN) to detect action concepts and adopted a linear dynamical system (LDS) for clip concepts relationship modeling.

Recently Tan and Yang\textsuperscript{19} proposed a general approach for a few shot action recognitions using feature aligning network (FAN) by applying alignment using CNN. Nazir et al.\textsuperscript{20} provide a better feature representation by combining the 3D Harris STIP detector and 3D scale-invariant feature by following traditional bag of visual feature approach and SVM classifier is used for training and testing. Camarena et al.\textsuperscript{21} explore the subject key points estimation by key trajectories which is eight-time faster comparable to the dense trajectories approach while Zho et al.\textsuperscript{22} proposed hierarchical temporal memory enhanced one-shot distance learning (HED).

Diman and Vishwakarma\textsuperscript{23} constructed a robust feature vector by computing R-transform and Zernike moments on average Silhouette images obtained from Microsoft’s Kinect sensor V1. A deep view-invariant-human action recognition framework using two-stream motion and shape temporal dynamics presented in Diman and Vishwakarma.\textsuperscript{24} While Vishwakarma and Singh\textsuperscript{25} proposed recently a framework using Key Pose frame selection by fuzzy logic and Pose feature extraction using Spatial Edge Distribution of Gradients (SEDGs) and Gabor filter. The next section covers the proposed schemes for action recognition.

**Proposed schemes**

We propose a hand-crafted human action recognition framework in which is going through six main processes to reach the highest recognition accuracy started by input video re-sampling, shot boundary detection, shot sub-sampling, feature extraction, compact feature vector representation, and classification as the final process. Three methods will be proposed where each one conducted in five steps before classification that common in the shot boundary detection and shot sub-sampling. The framework of each is depicted in Figure 1

\textbf{A. Common procedures:} Each of the three proposed approaches is common in three main steps, two mains at the beginning of the human action recognition process, and one ends the process, which we will explain as follows:

- **Shot boundary detection:** In this work, a simple shot boundary detection method has been used\textsuperscript{26} that is dependent on the difference of the luminance histogram along with an adaptive threshold for temporal subsampling. In shot boundary detection, the algorithm procedures involve four steps. The first step is frame rate resampling, where the input video is sampled at an interval of 0.2 s, which is considered a good compromise between
computational complexity reduction and visual information preservation. In the second step, the process involves distance measurement, where the measurement of a distance between $f_{(t+1)}$ and $f_t$ frame (after resampling) is evaluated by using the following formula:

$$d(t) = \frac{1}{N} \sum_{i=0}^{k} |\hat{h}_i(t) - \hat{h}_i(t + 1)|$$

where $\hat{h}_i(t + 1)$ and $\hat{h}_i(t)$ denote the aggregated luminance histogram for $f_{(t+1)}$ and $f_t$, respectively. $k$ denotes the number of bins for the histogram and $N$ denotes the number of pixels a frame is carrying. A large value of $k$ results in increased distance between two identical pixel values at dissimilar bins. Similarly, a too coarse value of $k$ will result in the loss of information, leading to an unsuccessful attempt to achieve the desired perceptive power. This will also result in disturbing the good balance between the two extremes. Thus, a relatively moderate bin number ($k = 128$) was used in the study. Third, the procedures involve the filtered distance measurement by

$$\tilde{d}(t) = |d(t) - MED(d(t-1), d(t), d(t + 1))|$$

Figure 1. Proposed framework of the three schemes.
where MED (.) denotes the “median filtering operation,” which helps to detect an abrupt illumination change. The fourth step is adaptive thresholding, that is, if $\tilde{d}(t)$ exceeds a certain threshold value, showing that the considered frame represented a unique incident and was represented as an “anchor frame,” where the threshold $T$ is calculated as follows:

$$T(t) = \mu_d(t) + \alpha \sigma_d^2(t)$$

(3)

In this equation, $\sigma_d^2(t)$ and $\mu(t)$ are the sample variance and sample mean of the filtered distance measured in window $W$ of length $L$ within a short time. Here, $\alpha$ is a scale factor. Mathematically, $\mu(t)$ and $\sigma_d^2(t)$ can be evaluated using the following relations:

$$\mu_d(t) = \frac{1}{L} \sum_{i \in w} \tilde{d}(i)$$

(4)

$$\sigma_d^2(t) = \frac{1}{L} \sum_{i \in w} (\tilde{d}(i) - \mu_d(t))^2$$

(5)

- Shot subsampling: Taking three frames that represent each shot; here, we take the first frame (the anchor frame), the middle frame, and the last frame of each shot.
- Classification: The final step is classification, that is, using different models, such as SVM, KNN, and Bagged Trees, with different kernel functions by picking only the most accurate results that are further compared with the results from the other schemes.

B. Core procedures:

(1) First scheme: It is conducted from four core steps which they are as follow:

(a) Get bag of words (BOW): This procedure involves treatment of each RGB frame as a document and defines the frame visual words; this is briefly achieved first through feature detection and description using a scale-invariant transform (SIFT) and second through codebook generation using K-mean clustering.

(b) Detect space-time interest points (STIP): This technique involves the detection of STIP of each grayscale frame. This is principally interesting because these frames concentrate on the information initially contained in many pixels on a few certain points that can further be related to the spatiotemporal occurrence in the sequence. This scheme typically helps in different joint motions, such as running, walking, or jumping.

(c) Apply the principal component analysis (PCA): The third and fourth steps described above concatenate the two output feature vectors. Afterwards, PCA is applied which shows the internal structure of the data in such a way that best declares the variance in data and for a lower-dimensional output vector.
(d) **Gaussian mixture model (GMM)**: This is applied to the final feature vector of each video to keep the initial characteristics of the extracted features and avoid distortion.

(2) **Second scheme**: The second scheme involves six procedures including the three core ones previously explained in the common procedures:

(a) **Feature extraction**: In this process, for each RGB frame, a histogram of oriented gradient (HOG) descriptors is extracted. The property of these descriptors is that they deliver excellent performance for human detection, which is increased by concatenation with the feature vector that was extracted by applying histogram of optical flow (HOF) on the same subsampled frames for robust motion detection.

(b) **Gaussian mixture model (GMM)**: This model is applied twice. It is first applied to each frame feature vector individually and then on the whole output feature vector of a given video. This application technique retains the initial characteristics of the extracted features and avoids distortion.

(c) **Applying fisher vector**: This is applied on reshaped clustered local descriptors to improve the classification performance on a fixed-length feature vector that can perform well even with simple linear classifiers.

(3) **Third scheme**: The third scheme consists of three steps in addition to the previous three mentioned common ones:

(a) **Image enhancement**: This involves the adjustment of each grayscale frame by normalization, that is, by decreasing the outcome of illumination differences using the following formula:

\[
y = \left( \frac{x - \text{MIN}}{\text{MAX} - \text{MIN}} \right) \cdot \frac{255}{\text{MIN}} \]

where “x” represents the grayscale value of the original image, “y” is the grayscale value of the output image after normalization, and MIN and MAX represent the minimum and maximum gray values of the original image, respectively.

Gamma correction frame adjustment is also applied in the simplest cases as a non-linear operation utilized to encode and decode luminance using the following power-law expression:

\[
v_{\text{out}} = A v_{\text{in}}^y
\]

In the above equation, if the gamma value “y” is less than 1, it will function as an encoding gamma. Conversely, if the gamma value y is greater than 1, it is a decoding gamma.

(b) **Binary robust invariant scalable key point (BRISK) detection**: BRISK is considered an efficient generation of key points and a method for the detection, description, and matching of image regions. This technique is also robust to all possible image transformations because of the binary nature of the BRISK descriptor.
In the present study, BRISK point detection is used on each frame. Afterwards, the strongest five points generated from each frame are selected.

(c) *Gaussian mixture model (GMM)*: GMM is applied to each video output feature vector to avoid feature distortion and retain its initial characteristics.

**Experimental results**

**Data sets**

The performance of the three proposed schemes was evaluated on four well-known databases and challenge human action datasets. A brief description of the employed datasets is given in the following sub-sections:

(1) *Hollywood2*\(^{27}\) contains 12 action classes collected from 68 different Hollywood movies. These action classes are “answering a phone,” “driving a car,” “hugging a person,” “fighting person,” “action of getting out of a car,” “kissing,” “handshake,” “run,” “sit down,” “eat,” “stand up,” and “sit up.” The action classes are present in 1707 videos. These videos were divided into two sets: 823 videos for the training set, and 844 for the testing set.

(2) *KTH*\(^{28}\) is comprised of six action classes defined from 600 black and white videos captured by a static camera. The action classes are “boxing,” “walking,” “jogging,” “running,” “hand waving,” and “clapping.” These videos were divided into two sets: 384 videos for the training set, and 216 videos for the testing set.

(3) *UCF11*\(^{6}\) contains 11 action classes: “soccer juggling,” “trampoline jumping,” “basketball shooting,” “volleyball spiking,” “horseback riding,” “walking with a dog,” “golf swinging,” “biking,” “swinging,” “tennis swinging,” and “diving.” These action classes are considered very interesting but challenging due to large variations in the appearance of the moving objects. Other factors that add to its complexity are camera motion, object scale, pose, lighting conditions, and mixed-up background.

(4) *Weizmann*\(^{29}\) is made up of 10 natural action groups conducted by nine individuals, such as Sprint, Hand, Skips, Run, Run-Jack, Bend, Jack, Walk, Wave-1 and Wave-2, and Wave-2. In total, there are 90 low-resolution video clips of 144 × 180 pixels.

**Experiments**

Experiments were conducted on the four datasets using the three proposed schemes. The experiment decreases the entire video frames, which reduces the burden on the processing and increases the accuracy of the results on three levels. In the first step, only one frame is selected (every 0.2 s) for shot detection in the multi-
shot datasets, such as Hollywood2. In the second step, three frames are taken representing each shot in the next scheme’s procedures applied on the Hollywood2, KTH, UCF11, and Weizmann datasets.

After the frame selection, various features, including BOW, STIP, HOG, and HOF, were extracted. The processing enhancements on each RGB or gray frame were also applied according to the feature specifications. Afterwards, various algorithms were applied such as GMM on a frame or a video output feature vector. This will enhance and reveal the internal structure of the represented output data due to application of the Expectation-Maximization (EM) method to reduce overfitting. Finally, Fisher vector and PCA were used to obtain the most compact fixed-length feature vector for good shot representation.

In the three proposed schemes, KNN, Bagged Trees, and SVM are used as classifiers of fivefolds cross-validation with different kernels selecting only the best results with their true positive \( (R) \), false negative \( (FN) \), false discovery rates \( (FD) \), and positive predictive values \( (P) \) which are calculated using the following formulas:

\[
R = \frac{TP}{TP + FN} \quad (8)
\]

\[
FN = 1 - R \quad (9)
\]

\[
P = \frac{TP}{TP + FP} \quad (10)
\]

\[
FD = \frac{1}{C0}R \quad (11)
\]

Figures 2 to 13 show the confusion matrix from the results of the classification of the three proposed schemes on the four datasets used. The true positive and false

![Figure 2. The first scheme on Hollywood2, Fine-KNN rates: (a) true positive and false negative rates and (b) positive predictive values and false discovery rates.](image-url)
negative rates are depicted in Figures 2(a) to 13(a) while Figures 2(b) to 13(b) show the positive predictive values and false discovery rates.

The aforementioned rates shown in Figures 2 to 5 are based on the KNN (fine and weighted kernels) and SVM (fine Gaussian kernel) classifiers after applying the first scheme on the Hollywood2, KTH, UCF11, and Weizmann datasets. Figures 6 to 12 show the confusion rates based on the KNN (weighted and subspace kernels) and Bagged Trees classifiers with the second and third schemes applied on the four datasets.

**Figure 3.** The first scheme on KTH Fine-KNN rates: (a) true positive and false negative rates and (b) positive predictive values and false discovery rates.

**Figure 4.** The first scheme on UCF11 fine Gaussian SVM rates: (a) true positive and false negative rates and (b) positive predictive values and false discovery rates.
It is observed that the positive predictive values are high in the verbs of the KTH, UCF11, and Weizmann but not quite high in Hollywood2 using the proposed three schemes where precision show how useful the search results are after applying each method on the datasets, the second scheme shows a high measure of quality in the precision values.

As shown in Table 1, the first scheme gives the accuracy results of 57.8%, 94.5%, and 98.9% using the KNN classifier applied on the Hollywood2, KTH, and Weizmann datasets, respectively, while it gives 94.5% accuracy result using SVM applied on the UCF11 dataset. The second scheme recorded the high
accuracy results of 73.6% and 97.0% using the KNN classifier on the Hollywood2 and KTH datasets, it gives 97.8% on Weizmann by SVM, respectively, and recorded 95.6% accuracy using Bagged Trees applied on the UCF11 data set. The third scheme gives accuracy results of 52.0% on Hollywood which is not high enough compared to 95.3% on KTH classified by KNN but recorded 94.2% and 100% accuracy result using Bagged Trees and SVM for classification on UCF11 and Weizmann. The recognition accuracy (RA) calculated according to the following formula:

$$RA = \frac{\sum TP + \sum TN}{T}$$  \hspace{1cm} (12)
From the results in Table 1, it can be seen that accuracy measures involving Hollywood2 dataset recorded lower achievements. This may be due to the nature of the Hollywood2 dataset which intends to provide a comprehensive benchmark for human action recognition in realistic and challenging settings like dynamic complex backgrounds, uncontrolled illumination conditions, and providing many variations in occlusions and scene surrounding. Despite the challenging datasets, the proposed schemes show good results. As shown in the table the KNN classifier shows its effect on the majority of high classification results compared to SVM then Bagged-Trees, that’s an effect due to the overcome on one of its disadvantages.

**Figure 9.** The second scheme on Weizmann weighted KNN rates: (a) true positive and false negative rates and (b) positive predictive values and false discovery rates.

**Figure 10.** The third scheme on Hollywood2 subspace-KNN rates: (a) true positive and false negative rates and (b) positive predictive values and false discovery rates.

From the results in Table 1, it can be seen that accuracy measures involving Hollywood2 dataset recorded lower achievements. This may be due to the nature of the Hollywood2 dataset which intends to provide a comprehensive benchmark for human action recognition in realistic and challenging settings like dynamic complex backgrounds, uncontrolled illumination conditions, and providing many variations in occlusions and scene surrounding. Despite the challenging datasets, the proposed schemes show good results. As shown in the table the KNN classifier shows its effect on the majority of high classification results compared to SVM then Bagged-Trees, that’s an effect due to the overcome on one of its disadvantages.
by computing and minimize the distances to all training objects by using GMM to only one neighbor.

Table 2 shows comparisons of the results obtained from the three proposed schemes to recent state-of-the-art techniques applied to the four datasets. It can be seen from the results that the first scheme achieved higher recognition accuracies than references:14,30 on Hollywood2,20,21,31,32 on KTH,33,34 on Weizmann by 2.9% and 1.4%. While overcoming all mentioned references on UCF11 except Gammulle et al.35 by 0.1% difference. The second scheme provides the highest recognition accuracies on all recently published references and recorded 97.0% on KTH, 95.6% on UCF11, 97.5% on Weizmann and 73.6% on Hollywood2, except Liu et al.32 by difference 1.4%.

Figure 11. The third scheme on KTH weighted-KNN rates: (a) true positive and false negative rates and (b) positive predictive values and false discovery rates.

Figure 12. The third scheme on UCF11 bagged trees rates: (a) true positive and false negative rates and (b) positive predictive values and false discovery rates.
The third scheme recorded higher recognition accuracies (RAs) only to the methods and Liu et al. on Hollywood2 by 52.0%, while it beats all the existing methods by on UCF11 by 94.2% except Gammulle et al. by 0.4% difference, and on Weizmann by 100% due to its simplicity and the effectiveness of the third approach. But the third scheme with the KTH gives completely unsatisfactory results compared to previously developed models due to the different ecological condition during recording the KTH which varying the illumination and environmental status where the image enhancements and BRISK points failed to handle. The second scheme showed a high achieved recognition accuracy on four various challenging datasets even with recently proposed approaches applied on Hollywood2, KTH, UCF11, and Weizmann. The first and third scheme its recognition accuracies vary according to the dataset’s challenging levels and conditions and how much it is compatible with the selected feature extraction methods.

**Figure 13.** The third scheme on Weizmann weighted KNN rates: (a) true positive and false negative rates and (b) positive predictive values and false discovery rates.

**Table 1.** Schemes accuracy results on the four datasets.

| Scheme | Hollywood2 Classifier | RA (%) | KTH Classifier | RA (%) | UCF11 Classifier | RA (%) | Weizmann Classifier | RA (%) |
|--------|------------------------|--------|----------------|--------|------------------|--------|----------------------|--------|
| 1      | KNN                    | 57.8   | KNN            | 94.5   | SVM              | 94.5   | KNN                  | 98.9   |
| 2      | KNN                    | 73.6   | KNN            | 97.0   | Bagged-trees     | 95.6   | SVM                  | 97.8   |
| 3      | KNN                    | 52.0   | KNN            | 95.3   | Bagged-trees     | 94.2   | SVM                  | 100    |
Table 2. Comparison of our results to those of recent state-of-the-art techniques on the three datasets.

| Hollywood2 | KTH | UCF11 | Weizmann |
|------------|-----|-------|----------|
| Method     | RA (%) | Method | RA (%) | Method | RA (%) | Method | RA (%) |
| R.T \textsuperscript{30} | 45.7 | P-CNN \textsuperscript{31} | 90.0 | CNNs + LDS \textsuperscript{18} | 86.16 | GWT + RT \textsuperscript{33} | 96.0 |
| G.P \textsuperscript{14} | 46.8 | G.P \textsuperscript{14} | 95.8 | G.P \textsuperscript{14} | 82.3 | AEIs + SDGs \textsuperscript{34} | 97.5 |
| R-NKTM \textsuperscript{36} | 59.4 | sDTD \textsuperscript{17} | 96.8 | LSTM \textsuperscript{35} | 94.6 | | |
| HC-MTL \textsuperscript{32} | 75.0 | HC-MTL \textsuperscript{32} | 94.3 | HC-MTL \textsuperscript{32} | 64.0 | | |
| TSVW \textsuperscript{15} | 61.3 | TSVW \textsuperscript{15} | 95.1 | TMNF + IVLAD \textsuperscript{37} | 91.4 | | |
| FVEM \textsuperscript{13} | 65.6 | Ips + FS \textsuperscript{16} | 95.8 | LCGSC \textsuperscript{38} | 69.5 | | |
| | | 3D-STVM \textsuperscript{19} | 95.8 | | | | |
| | | Key Traj \textsuperscript{21} | 94.20 | Key Traj \textsuperscript{21} | 60.3 ± 0.1 | | |
| | | 3D Harris + 3D SIFT \textsuperscript{20} | 91.82 | FAN \textsuperscript{19} | 80.66 | | |
| | | GWT + RT \textsuperscript{33} | 96.66 | HED \textsuperscript{22} | 60.3 | | |
| | | AEIs + SDGs \textsuperscript{34} | 96.6 | | | | |
| First scheme | 57.8 | First scheme | 94.5 | First scheme | 94.5 | First scheme | 98.9 |
| Second scheme | 73.6 | Second scheme | 97.0 | Second scheme | 95.6 | Second scheme | 97.8 |
| Third scheme | 52.0 | Third scheme | 59.3 | Third scheme | 94.2 | Third scheme | 100 |


Overall performance comparison

As an overall evaluation, our three approaches obtain competitive recognition accuracy results but especially in the second scheme as shown in Table 2 on different challenging, realistic datasets including Hollywood2, KTH, UCF11, and Weizmann. The excellent performance is attributable to minimising the number of frames representing each video and robust feature values that efficiently discriminate among patterns of different classes such as HOG used in the second scheme with Dalal-Triggs variant, $4 \times 4$ cell size, and HOF by the Lucas-Kanade method with a 0.009 threshold for noise reduction.

We shall continue with efficient regularization and training using the GMM used in the three schemes with the expectation-maximization (EM) method of 50 clusters to construct a visual word dictionary that retains the strongest patterns of the frame and video representations. Fisher vector used in the second scheme on the GMM video output improves the classification performance on the representation by using non-linear additive Hellinger’s kernel and normalization.

The results showed that KNN with different kernels: for example, a subspace kernel of 30 random subspace learners and 50 dimensions, the weighted kernel of 10 neighbors and squared inverse distance weight, the fine kernel of 1 neighbor, and Euclidean distance metric is the best classifier on Hollywood2, KTH, and Weizmann in case of the three schemes; whereas Bagged Tree using Random Forest and Decision Tree is the best on UCF11 in case of the second and third schemes. SVM with fine Gaussian kernel is best in case of the first scheme. Thus, our approaches are considered worthy schemes as they give excellent results compared to other benchmarked approaches.

Conclusions and future work

This paper presents three different schemes for action recognition by extracting various local and holistic features trying to emphasize different aspects of actions that are suitable for different datasets. The feature selection performs fusion and enhanced the performance by obtaining an optimal feature vector as a final video representation while considering the effect of the feature vector length on the recognition accuracy and the different models, including KNN, Bagged Tree, and SVM, used for classification.

We tried to choose the most relevant features and discard the irrelevant ones as the accuracy of the learned function depends strongly on how the input objects represented for constructing a simple model to make features easier to interpret, for shorter training time by avoiding the curse of dimensionality but containing enough information. Enhanced generalization occurs by reducing overfitting or reduction of variance using GMM. Experimental results show that each video is short-time processed and compact feature vector length using shot boundary detection also due to the re-sampling frames representing each shot. Finally, a compact fixed feature vector length representing a video of the concatenated different extracted
features. Each of the proposed schemes provides high recognition accuracy compared to other state-of-art methods.

As an overall conclusion, our approaches get a highly competitive recognition accuracy results on different challenging realistic datasets like Hollywood2, KTH, UCF11, and Weizmann used by recent state of art methods. Especially, the Second Scheme so it considered a worthy Scheme with the Subspace or weighted-KNN classifier as it gives excellent results compared to other benchmarked approaches.

Until today the human action recognition research field limitation is still not reaching the optimum recognition accuracy by one common algorithm for available datasets which represent the constraint we face in challenging real world. In the future, this study can be extended to extract other features that enhance the recognition accuracy and to handle more challenging videos to be more effective in the action recognition task.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Cherry Aly https://orcid.org/0000-0002-8744-8793

References
1. 37 Mind-Blowing YouTube Facts, Figures and Statistics – 2019, https://merchdope.com/youtube-stats/ (accessed 1 February 2020).
2. Subetha T and Chitrakala S. A survey on human activity recognition from videos. In: International conference on information communication and embedded systems (ICICES), Chennai, 25–26 February, 2016. Chennai:IEEE.
3. Dollar P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse Spatio-temporal features. In: International workshop on visual surveillance and performance evaluation of tracking and surveillance, Beijing, China, 15–16 October 2005. Beijing:IEEE.
4. Laptev I. On space-time interest points. Int J Comp Vis 2005; 64(2): 107–123.
5. Scovanner P, Ali S and Shah M. A 3-dimensional SIFT descriptor and its application to action recognition. In: Proceedings of the 15th international conference on multimedia, Augsburg, Germany, 2007, pp.357–360.
6. Liu J, Luo J and Shah M. Recognizing realistic actions from videos in the wild. In: 2009 IEEE conference computer vision and pattern recognition, Miami, FL, 20–25 June 2009, pp.1–8. Florida:IEEE.
7. Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies. In: *IEEE conference on computer vision and pattern recognition*, Anchorage, AK, 23–28 June 2008, pp.1–8. Alaska:IEEE.

8. Klaser A and Marszalek M. A spatio-temporal descriptor based on 3D-gradients. In: *Proceedings of the British machine vision conference*, Leeds, 2008, pp. 995–1004.

9. Bobick A and Davis J. The recognition of human movement using temporal templates. *IEEE Trans Pattern Anal Mach Intell* 2001; 23(3): 257–267.

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

11. Han J, He S, Qian X, et al. An object-oriented visual saliency detection framework based on sparse coding representations. *IEEE Trans Circ Syst Video Technol* 2013; 23(12): 2009–2021.

12. Junwei H, Xiang J, Xintao H, et al. Representing and retrieving video shots in human-centric brain imaging space. *IEEE Trans Image Process* 2013; 22(7): 2723–2736.

13. Sekma M, Mejdoub M and Amar C. Structured fisher vector encoding method for human action recognition. In: *15th International conference on intelligent systems design and applications, ISDA 2015*, Marrakech, Morocco, 2015, pp.14–16. Marrakech:IEEE.

14. Liu L, Shao L, Li X, et al. Learning spatio-temporal representations for action recognition: a genetic programming approach. *IEEE Trans Cybern* 2016; 46(1): 158–170.

15. Kardaris N, Pitsikalis V, Mavroudi E, et al. Introducing temporal order of dominant visual word sub-sequences for human action recognition. In: *IEEE international conference on image processing (ICIP)*, Phoenix, AZ, 2016, pp.3061–3065. Phoenix:IEEE.

16. Xu K, Jiang X and Sun T. Two-stream dictionary learning architecture for action recognition. *IEEE Trans Circ Syst Video Technol* 2017; 27(3): 567–576.

17. Shi Y, Tian Y, Wang Y, et al. Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Transactions on Multimedia* 2017; 19(7): 1510–1520.

18. Zhang L, Feng Y, Xiang X, et al. Realistic human action recognition: when CNNS meet LDS. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp.1622–1626.

19. Tan S and Yang R. Learning similarity: feature-aligning network for few-shot action recognition. In: *International joint conference on neural networks*, Budapest, Hungary, 2019. Pudapest:IEEE.

20. Nazir S, Yousaf M and Velastin S. Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition. *Comput Electr Eng* 2018; 72: 660–669.

21. Camarena F, Chang L and Gonzalez-Mendoza M. Improving the dense trajectories approach towards efficient recognition of simple human activities. In: *2019 7th International workshop on biometrics and forensics (IWBF)*, Cancun, Mexico, 2019. Pudapest:IEEE.

22. Zou Y, Shi Y, Wang Y, et al. Hierarchical temporal memory enhanced one-shot distance learning for action recognition. In: *IEEE international conference on multimedia and expo (ICME)*, San Diego, CA, 2018. San Diego:IEEE.

23. Dhiman C and Vishwakarma DK. A robust framework for abnormal human action. *IEEE Sens J* 2019; 19(15): 5915–5203.

24. Dhiman C and Vishwakarma DK. View-invariant deep architecture for human. *IEEE Trans Image Process* 2020; 29: 3835–3844.
Appendix

Nomenclature

\( N \) number of pixels a frame is carrying

\( \hat{h}_t(f) \) aggregated luminance histogram for \( f \)

\( k \) number of bins for the histogram

\( \mu(t) \) sample mean
x  grayscale value of the original image
\(y\)  grayscale value of the output image after normalization
\(v_{\text{out}}\)  output pixel value
\(v_{\text{in}}\)  input pixel value
\(A\)  positive constant
\(RA\)  recognition accuracy
\(R\)  recall (true positive rates)
\(P\)  precision (positive predictive values)
\(FN\)  false negative rates
\(FD\)  false discovery rates
\(TP\)  correctly identified patterns
\(FP\)  incorrectly identified patterns
\(TN\)  correctly rejected patterns
\(FN\)  incorrectly rejected patterns

**Greek letters**

\(\alpha\)  scale factor

**Author biographies**

Cherry A. Aly received the M.S. degree in computer engineering from Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt. Working as a lecturer and currently pursuing her PhD degree in computer engineering at Multimedia University, Faculty of Engineering and Technology, Melaka, Malaysia. Her research interest covers Video-Image Processing and Analysis, Pattern Recognition and Computer Vision.

Fazly S. Abas obtained his PhD from the University of Southampton, United Kingdom in 2004 in the field of image analysis and pattern recognition. Currently, he is a dean of the Faculty of Engineering and Technology at Multimedia University, Malaysia. His research interests include Image Processing and Analysis, Digital Pathology, Medical Image Analysis and Video Analytics.

Goh H. Ann received the B.S. and M.S. degree in engineering from Universiti Teknologi Malaysia. He obtained his PhD degree From Multimedia University, Malaysia. Since then, has been working as a senior lecturer in MMU, Malaysia. His research interest includes (but not limited to) Image Processing and Analysis, Computational Intelligence and Data Mining.