IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 June 2023, accepted 14 July 2023, date of publication 17 July 2023, date of current version 24 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296252

RESEARCH ARTICLE

ViCo-MoCo-DL: Video Coding and Motion Compensation Solutions for Human Activity Recognition Using Deep Learning

TAMER SHANABLEH, (Senior Member, IEEE)
Department of Computer Science and Engineering, American University of Sharjah, Sharjah, United Arab Emirates
e-mail: tshanableh@aus.edu

This work was supported in part by the Open Access Program through the American University of Sharjah under Grant OAPCEN-1410-E00189.

ABSTRACT This paper proposes three novel feature extraction approaches for human activity recognition in videos. The proposed solutions are based on video coding concepts including motion compensations and coding based feature variables. We use these features with deep learning for model generation and classification, hence the ViCo-MoCo-DL abbreviation which stands for Video Coding and Motion Compensation with Deep Learning. These solutions are fused in terms of averaging their classification scores to predict the human activity in videos. In all proposed solutions, an input video is temporarily segmented into 12 non-overlapping segments of equal size. In the first and second solution each segment is converted into one component of an RGB image, thus resulting in 4 RGB images. The conversion happens in terms of motion capture using motion estimate, motion compensation and accumulating image prediction errors. Consequently, in the first solution, the 4 generated RGB images are tiled into one big image which is used to train a Convolutional Neural Network (CNN) network. In the second solution each generated RGB image is entered into a pre-trained CNN for feature extraction. The resultant FVs are arranged into a matrix and used for training a Long Short-Term Memory network (LSTM). In the third solution, a customized High Efficiency Video Coder (HEVC) is used to generate feature variables per frame. The resultant Feature Vectors (FVs) of 3 video segments are arranged into a matrix and numerically summarized into one FV, thus, each input video is represented by 4 FVs which are used to train another LSTM network. Experimental results on three well-known datasets show the superior classification results of the proposed fused solution over existing work.

INDEX TERMS Deep learning, human action recognition, video classification, motion compensation, video coding.

I. INTRODUCTION

Human Activity Recognition (HAR) is at the heart of human-computer interaction applications. In general, HAR applications include healthcare, surveillance and entertainment. The HAR applications in healthcare include gait analysis, fitness and monitoring patients and the elderly. Additionally, the HAR applications in surveillance include monitoring in smart homes and intruder detection and lastly, the HAR applications in entertainment include gesture recognition in gaming and remote device control.

Human Activity can be recognized with employing a variety of technical solutions include wearable sensors [1] and [2], smart phones [3] and [4], depth cameras [5], human skeleton estimation [6] and [7] and video-based activity recognition [8], [9], and [10].

This paper focuses on human activity recognition based on recorded videos with the use of deep learning. Such approaches are reported extensively in literature, for instance, in [11] temporal image representations of videos based on rank-pooling are used with pre-trained CNNs for

The associate editor coordinating the review of this manuscript and approving it for publication was Yun Zhang.

activity recognition. In [12] it was proposed to combine spatial and temporal context in one self-supervised framework and divide video frames into grids of patches, which are used to train a deep learning network. In [13] it was proposed to extend existing video datasets by applying action recognition on millions of videos over social platforms and categorizing them into existing categories of well-known datasets. The work reported in [13] was then extended by [14] to encode spatio-temporal feature variables where 2D convolution is performed along various views of video to learn temporal motion cues and spatial appearances. In [15], it was proposed to use Visual Geometry Group-19 (VGG19) along with multi-view features computed from vertical and horizontal gradients. Consequently, the best features are used for activity recognition.

In [16] a unique training was proposed in which a pre-trained CNN is used for the spatial stream on single frames of an input video. Whereas for the temporal stream, three different RGB frames are selected, converted into grayscale and combined into one RGB image. These images are then used with a pre-trained CNN for training and classification. In [17], motion representations are improved by tweaking the spatial stream to predict the outputs of the temporal stream and thus removing the need for a two-stream architecture. In [18], the authors used dilated CNNs to extract salient feature variables and fed them to an LSTM network. This is followed by an attention mechanism to enhance the classification accuracy. In [19], the authors proposed to combine spatial and temporal feature variables into one set of spatio-temporal features and employed the OFF CNN architecture for feature extraction.

In a recent work reported in [20], optical flow images between consecutive images are calculated and represented by a pre-trained CNN network. Principle Component Analysis (PCA) is then used to reduce the dimensionality resulting in a time series that can be represented and classified by a multi-channel 1D-CNN.

Many video-based feature extraction methods are reported in the literature for activity recognition. For instance in [21] the input video is median filtered followed and segmented using watershed and feature extraction including Histogram of Oriented Gradient, GiST and Color and a fusion of all Features. Likewise, feature extraction based on a fusion of discrete wavelet transformation, multiscale local binary patterns and histogram of oriented gradients are reported in [22]. In [23], motion is extracted from video and used as feature vectors based on a novel feature that relies on differential motion descriptors.

In this work on the other hand, video coding techniques are used for feature extraction including motion compensation and feature variables generated from coding an input video using High Efficiency Video Coding (HEVC) [24]. Three solutions are proposed and then fused at a classification score level to recognize the human activity in a given video. In one solution, motion is captured using motion vectors and motion compensation, consequently 4 RGB images are constructed using the motion information. The four images are tiled to generate one big image that is inputted to a CNN for training and classification. In the second solution, each of the constructed 4 RGB images that capture the motion go through a pre-trained CNN to generate feature vectors. These feature vectors are then stacked to generate a feature matrix that can be used with an LSTM classifier. The third solution uses a HEVC video coder to generate feature variables and then use statistical summaries to form feature vectors that are stacked to generate a feature matrix suitable for LSTM networks.

The rest of the paper is organized as follows. Section II introduces the three proposed solutions listed above. Section III introduces the proposed motion capture solutions. Section IV details the score fusion technique used; Section V reviews the datasets used for training and classification; Section VI presents the experimental results and Section VII concludes the paper.

II. PROPOSED SOLUTION

In this work we propose a classification solution composed of three parallel feature extraction and classification pipelines based on spatial features, temporal features and HEVC video coding features which are considered spatio-temporal features. The solutions are fused at a class score level by means of averaging. The three proposed solutions are independent solutions, the output of one solution is not inputted into another solution rather, the output classification scores of the three solutions are fused as a post-process.

Figure 1 presents the overall flowchart of the proposed training system. The three proposed solutions can extract features and build classification models in parallel. The three generated models are then saved for use in classifying test video sequences.

Figure 2 presents the overall flowchart of the proposed recognition of human activity. The three proposed solutions can classify the human activity in parallel. This is followed by class-score fusion to generate the final label of the human activity in the video.

As illustrated in both figures, each of the proposed solutions is a standalone classification system that contains feature extraction, model generation and classification.

The three proposed pipelines are introduced in details in the following subsections.

A. PROPOSED SPATIAL FEATURES AND CLASSIFICATION

In this proposed solution, the whole input video is represented as one RGB image that captures the motion traces of the video. The video is first temporally segmented into 12 equal size non-overlapping segments, then the motion in each segment is captured using either accumulated image differences or image differences with motion compensation as explained in details in Section III below. The captured motion is represented as one 2D array. Hence, capturing the motion of three video segments results in three 2D arrays that can be combined into one RGB image. Using this approach, capturing the motion of 12 video segments results in four

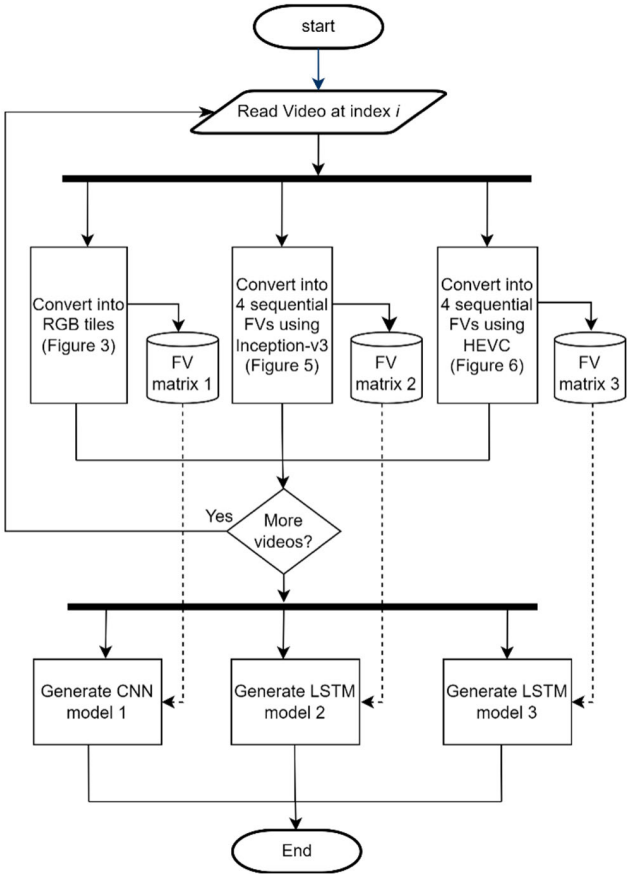

FIGURE 1. Overall flowchart of the proposed training system.

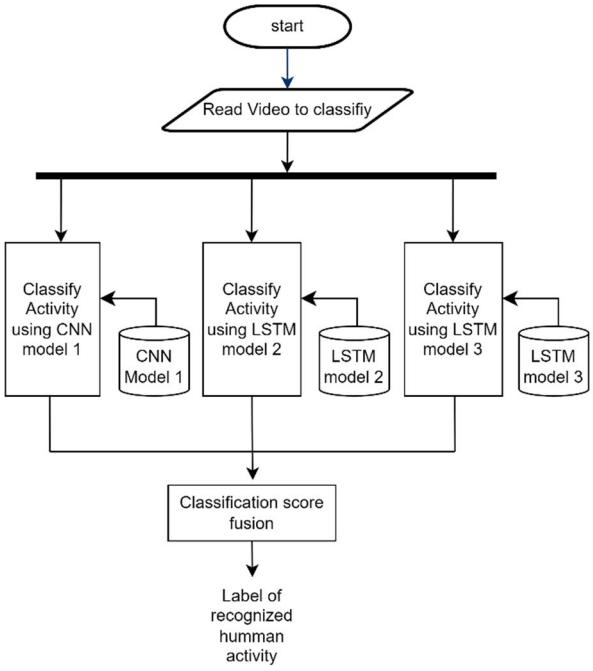

FIGURE 2. Overall flowchart of the proposed recognition of human activity.

RGB images, these images are then arranged into a square of 2 × 2 images and stored as one big RGB image. This proposed solution is further illustrated in Figure 3.

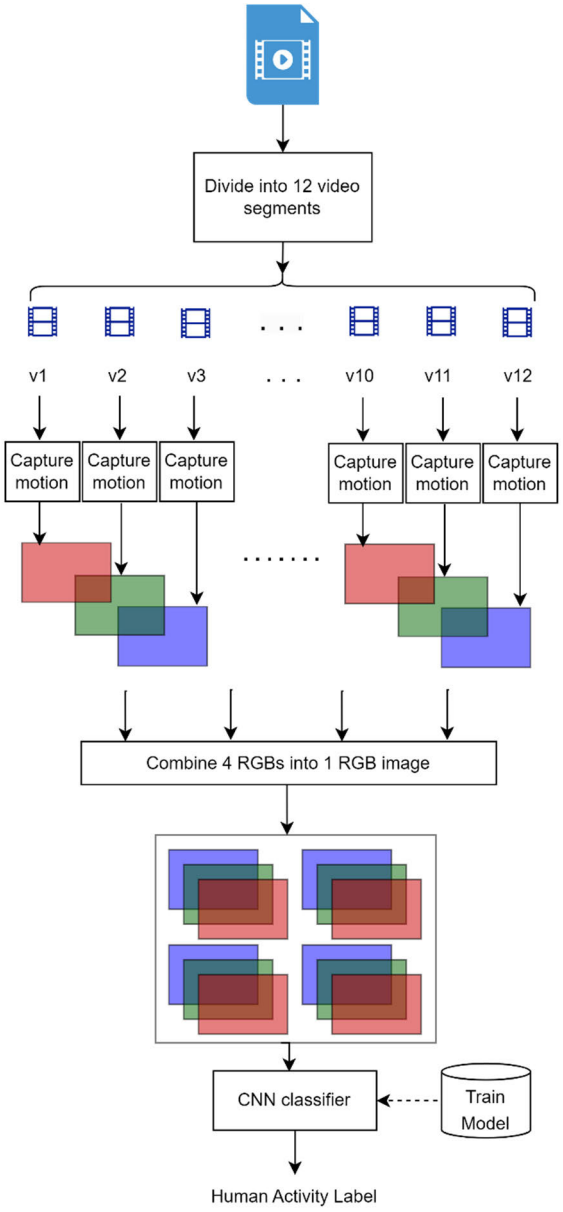

FIGURE 3. Proposed solution of converting a video into one RGB image and using CNN classification.

Once the video is represented as one RGB image, the human activity can be classified using any of the well-known CNN networks. In this work, we trained the Inception-v3 CNNs using the Joint-annotated Human Motion Data Base (jHMDB), HMDB51 and UCF11 datasets. The choice of Inception-v3 was empirical as it generated the best classification results. Additionally, the choice of 12 video segments is not arbitrary as such a number of segments allows for generating 4 RGB images that can be tiled or arranged into one big image which can be used as an input to CNN networks.

Example images resulting from this proposed solution are shown in Figure 4.

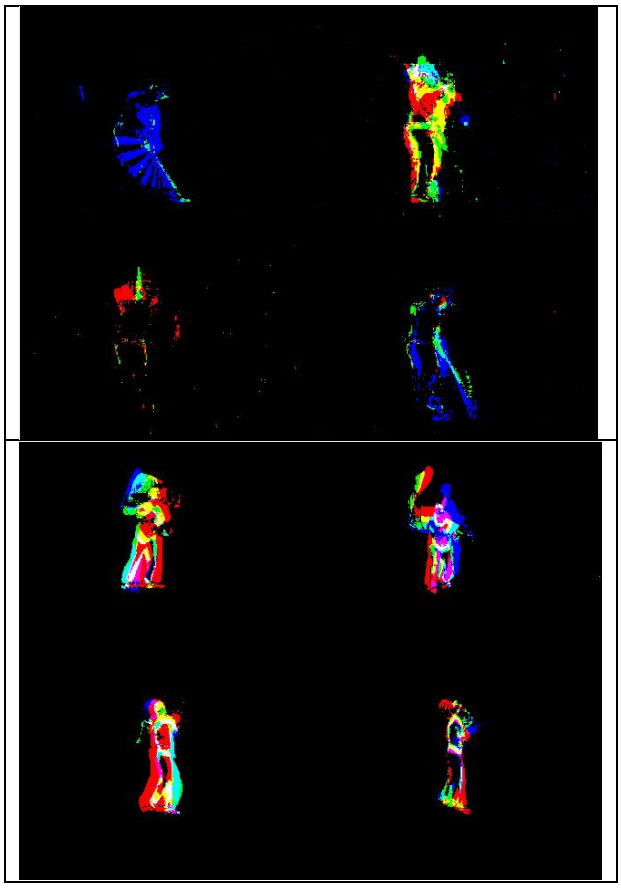

FIGURE 4. Example images resulting from this proposed solution of Figure 3. The examples are form the UCF11 dataset of classes Golf and Tennis. Each 2 × 2 images are used as one big RGB image for training and classification.

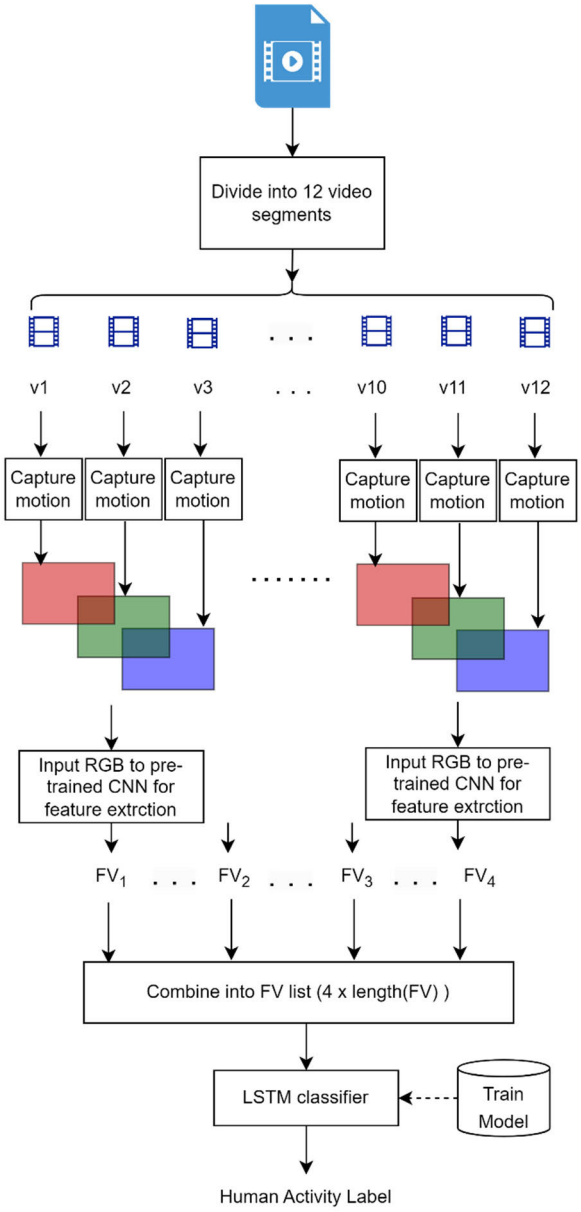

FIGURE 5. Proposed solution of converting a video into a list of feature vectors using motion capture and pre-trained CNN features.

B. PROPOSED TEMPORAL FEATURES AND CLASSIFICATION

Similar to the previous proposed solution, the video is temporally segmented into 12 equal size non-overlapping segments. Then, the motion in each segment is captured using either accumulated image differences or image differences with motion compensation as explained in details in Section III below. The captured motion is represented as one 2D array and the captured motion of three video segments are represented as one RGB image. Consequently, with 12 video segments, 4 RGB images are generated.

Each of these 4 RGB images is passed through a pre-trained Inception-v3 CNN network for feature extraction. The features are the output of the last 'avg_pool' layer and has a length of 2048 variables. Again, the choice of Inception-v3 is empirical as it generated the best results in this work. Since we have 4 RGB images, in this solution 4 feature vectors are generated. These feature vectors are time-wise sequential as the 12 video segments are sequential as well. Hence, they can be used with an LSTM classifier as illustrated in Figure 4.

Once, the video is represented as 4 sequential feature vectors, the human activity can be classified using an LSTM network. Again, in this work, we trained LSTM networks using the jHMDB, HMDB51 and UCF11 datasets. The architecture of the LSTM network contains an LSTM layer with 2000 nodes, followed by 50% dropout layer, these 2 layers are replicated followed by a fully connected softmax and a classification layers.

C. PROPOSED SPATIO-TEMPORAL HEVC FEATURES AND CLASSIFICATION

The HEVC video coding standard [24] is a revolution in video compression in that it achieves the same video quality as previous coding standards at a considerably lower bitrate.

Likewise, it generates videos with considerably higher quality than previous coding standards at the same bit rate.

The author of this work previously proposed extracting video features from the HEVC encoding process for use in intelligent video-based systems. This includes identifying key frames [25], prediction-based video scrambling [26] and encryption [27], detecting double and triple video compression [28], detecting motion vector data embedding in videos [29] and predicting video compression modes [30] to mention but a few.

The reason that concise and precise features can be extracted from HEVC-videos is that the mentioned video coder divides an input frame into large blocks known as Largest Coding Units (LCUs). These LCUs are then split recursively into various block sizes ranging from 64 × 64 to 8 × 8 pixels. The split is based on the spaio-temporal activity of the underlying video frame. These blocks are further divided into Prediction Units (PUs) with sizes ranging from 64 × 64 to 4 × 4. These PUs are used to calculate prediction error using motion estimation and compensation. To code the prediction error, a block is further divided into Transform Units (TUs) with block sizes ranging from 32 × 32 to 4 × 4 pixels.

In this work, we use HEVC-based video features for the purpose of recognizing human activity in videos. This includes 64 features based on the recursive splitting of LCUs, motion vectors, prediction errors, intra, inter and skipped coding modes, count of coding bits and so forth. Again, the use of similar features has proven successful in many other application as mentioned above, these feature are listed in Table 1.

TABLE 1. HEVC feature set used for human activity recognition.

ID	Feature variable (average per frame)
1	Number of CU parts
2	MVD bits per CU
3	CU bits excluding MVD bits
4	Percentage of intra CU parts
5	Percentage of skipped CU parts
6	Number of CUs with size 64x64
7	Number of parts with size 32x32
8	Number of CUs with size 16x16
9	Number of CUs with size 8x8
10	Row-wise SAD of the CU prediction error
11	Column-wise SAD of the CU prediction error
12	Ratio of gradients per CU
13	Total distortion per CU
Feature variable (one per frame)	
14-23	Standard deviation of feature IDs 1-9
24	Smallest CU size per frame
25-28	Standard deviation of feature IDs 10-13 per frame
29	Variance sum of the x and y components of frame MVs
30	For CUs with sizes > 8x8, $\log_2$(|sum of MV Differences|)
31	For CUs with sizes of 8x8, $\log_2$(|sum of MV DDifferences|)
32-48	Histogram of x-component of frame's MVs
49 -64	Histogram of y-component of frame's MVs

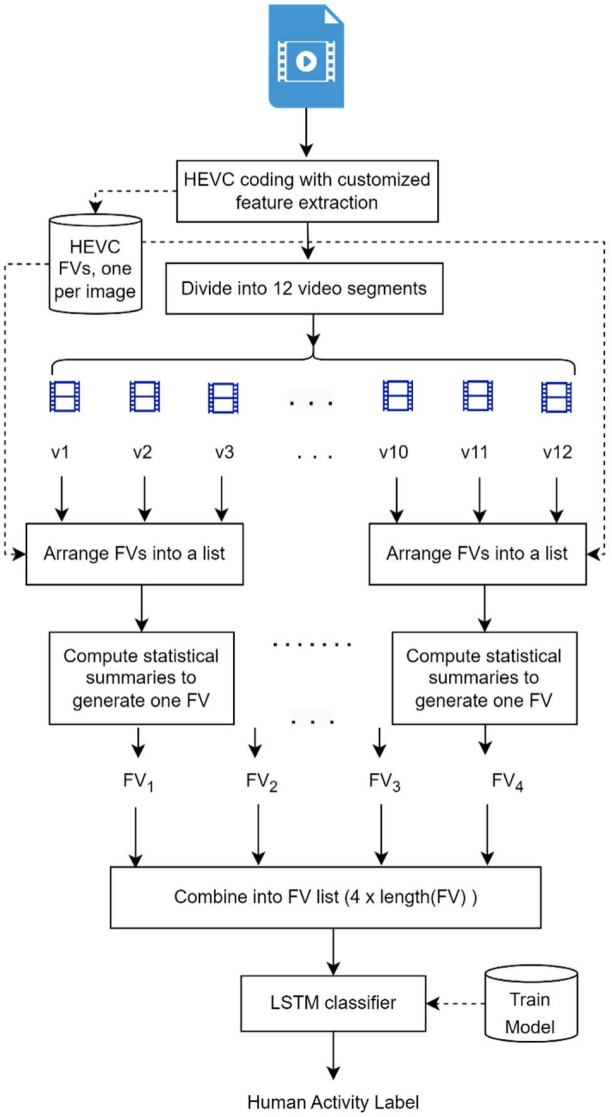

FIGURE 6. Proposed solution of converting a video into a list of feature vectors using HEVC features and statistical summaries.

To be consistent with the proposed solutions presented in Figures 1 and 3 above, in this solution we use the arrangement presented in Figure 4 to convert a video sequence into a list of FVs.

At first, an input video is compressed using a customized HEVC encoder that outputs the feature variables listed in Table 1. Consequently, the input video is temporally segmented into 12 equal size non-overlapping segments, and the FVs of each 3 segments are arranged into a list of FVs or FV matrix. Statistical measures are then applied to this FV matrix to summarize it into one FV. This results in 4 FVs per video sequence which is consistent with the proposed solutions presented in Figures 3 and 5 above. This arrangement is unique to this work and has not been reported previously.

The statistical summaries used in this work include the following for each HEVC feature variable: mean, standard

deviation, sum of absolute differences, mean of differences, standard deviation of differences and Entropy of differences. Such statistical measures have been used successfully by the author in the domain of classifying cognitive workload levels of as reported in [31].

III. MOTION CAPTURE

The proposed solutions illustrated in Figures 3 and 5 above, contain a process labeled as ''capture motion''. In this section we present the details of the proposed motion capturing used in our solutions.

More specifically, we propose two approaches for motion capture, one is based on our previous work on sign language recognition, which is known as accumulated image differences [32] and the other is based on a video coding concept known as motion compensation.

A. ACCUMULATED IMAGE DIFFERENCES

In this solution, absolute image differences are thresholded, binarized and accumulated into one image. The threshold is calculated empirically and in this work it is set to the third quartile of each image difference. Mathematically the thresholds of the image differences are represented as follows:

$$TH^i = sort(diff^i)_{q_3^i}, \quad i = 1..N-1 \tag{1}$$

where:

$$diff^i = \left| img^{i+1} - img^i \right|, \quad i = 1..N-1 \tag{2}$$

$$q_3^i = {}^3\!/_4 * length(diff^i), \quad i = 1..N-1 \tag{3}$$

where N is the total number of images and q_3 stands for the index of the third quartile of the image difference. Consequently, the Accumulated Differences Image (ADI) is represented as:

$$ADI = \sum_{i=1}^{N-1} \sum_{j=1}^{wxh} \begin{cases} 1, & diff_j^i > TH^i \\ 0, & diff_j^i \leq TH^i \end{cases} \tag{4}$$

where w and h are image width and height respectively.

B. IMAGE DIFFERENCES WITH MOTION COMPENSATION

Motion compensation is a technique used in video coding, it uses the motion vectors generated by the motion estimation process between two images to align the best match locations of an anchor image to the corresponding coordinates of the current image. Consequently, the current image is subtracted from the resultant motion compensated image to generate what is known as the prediction error [33].

In this work, we use optical flow to generate motion vectors between consecutive images. These motion vectors are used for motion compensation prior to subtracting images and accumulating their differences. Mathematically, the image difference with motion compensation is represented as:

$$diff_{x,y}^i = img^{i+1}(x,y) - img^i\left(x - Vx_{x,y}^i, y - Vy_{x,y}^i\right)$$
$$i = 1..N-1, \quad x = 1..w, \; y = 1..h \tag{5}$$

FIGURE 7. Example images to show the difference between the two motion capture approaches. Videos are from the Horse Riding class of the UCF11 dataset. (a) With Accumulated image differences (b) with accumulated image differences using motion compensation.

where Vx^i *and* Vy^i are the x and y motion vector components of i^{th} image and w and h are image width and height respectively.

Consequently, the Accumulated Differences Image with Motion Compensation (ADI_MC) is represented as:

$$ADI = \sum_{i=1}^{N-1} diff^i \tag{6}$$

Lastly, figure 5 presents example images to show the difference between the two motion capture approaches.

As shown in the figure, the two motion capture solutions results in different output images, as expected the output of the image differences with motion compensation is more refined as the subtraction and accumulation is applied after motion compensation. I can be argued though, that the first approach where motion is captured with accumulated image differences generates more information. In the Experimental results it is shown that both solutions complement each other and both are needed for fusing the classification scores.

IV. FUSION OF THE PROPOSED SOLUTIONS

In general, fusion can be done at a feature level where FVs of various solutions are concatenated. Fusion can also be applied at a score level, where classification scores of various

solutions are fused in terms of average, maximum or multiplication of scores.

In this work, we found that the best fusion approach is at a score level by means of computing the average classification scores of the proposed solutions followed by the prediction of the class label. Once the classification scores are averaged over the three proposed solutions, the maximum averaged score is selected and the corresponding class label is outputted. This arrangement is illustrated in Figure 8.

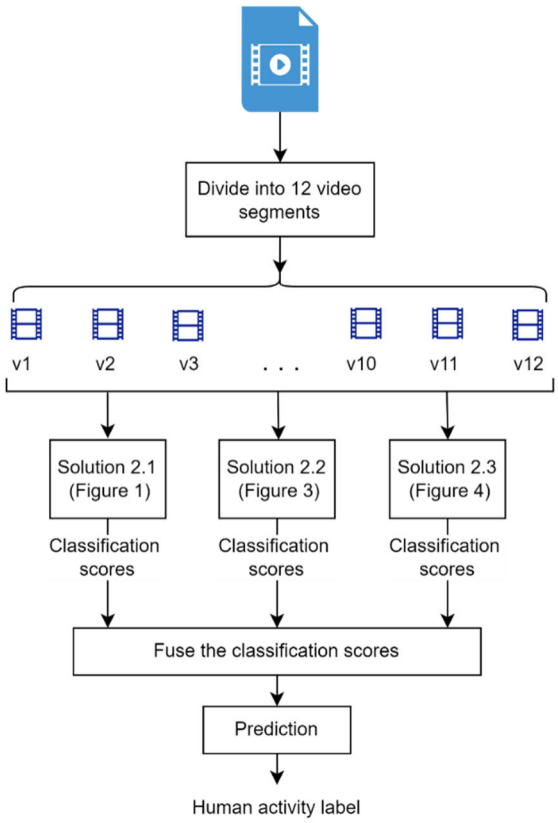

FIGURE 8. Score level fusion of the proposed solution.

In the experimental results it is shown that such a fusion approach worked well, as the three proposed solutions cover the spatial, temporal and spatio-temporal aspects of the human activity.

V. DATASETS

The solutions proposed in this paper are compared against existing work that depend on deep learning. Three commonly used datasets are employed in the experimental result; namely, HMDB51, JHMDB and UCF11.

The HMDB dataset [34] consists of 51 activity classes with a total of 6766 videos. The videos contain plenty of camera motion which is not necessarily centered on people and the speed of body parts vary across the videos. This dataset has 3 splits and the evaluation is based on the average classification results of the 3 splits.

The JHMDB dataset [35] is a subset of HMDB dataset and consists of 21 activity classes with a total of 923 videos from which around 70% is allocated for training and the rest for testing. The fewer number of videos in this dataset makes it challenging for training and classification.

Lastly, the UCF11 dataset [36] consists of 11 sport classes of humans like biking, volleyball and the like. The dataset has 1600 videos collected from YouTube and training and testing for this dataset follows a leave-one-out strategy. Among the three datasets used, this is the simplest dataset to classify.

VI. EXPERIMENTAL RESULTS

In the results to follow, the proposed solutions of Sections II-A, II-B and II-C are referred as CNN Tiles, CNN FVs and HEVC FVs respectively. Note that as mentioned in Section III, two solutions are proposed for motion capture, we experiment with both these solutions with the CNN FVs approach of Section II-B. Thus, two sub solutions are used; CNN FVs Diff and CNN FVs MC, where Diff stands for image difference and MC stands for motion compensation.

We start by presenting the classification results of each of the proposed solutions for the three datasets, jHMDB, HMDB51 and UCF11. The results in Table 2 present both the individual and fused classification results of each solution. All presented results are the average classification accuracies of the data splits.

TABLE 2. Classification results of all proposed solutions and final classification results using score fusion.

Proposed Solution	jHMDB	HMDB51	UCF11
CNN Tiles: Section 2.1	66.3%	51.8%	86.82%
CNN FVs MC: Section 2.2	66.4%	48.9%	82.64%
CNN FVs Diff: Section 2.2	56.0%	42.4%	80.1%
HEVC FVs: Section 2.3	58.9%	44.0%	88.2%
Score Fusion	**78.5%**	**71.41%**	**97.0%**

A number of observations can be drawn from the table as follows. First, the best classification result of the proposed solutions of Sections II-A, II-B and II-C differs from one dataset to the other, hence there is no one best solution. Second, the proposed solution of motion capture with motion compensation (Section III-B) consistently results in higher classification accuracy in comparison to the motion capture with accumulated image differences (Section III-A). This indicates that the novel approach of using motion compensation in motion capture works well.

The third observation drawn from the results in Table 2 is the most important of all. The score level fusion approach of the proposed solution produces excellent classification accuracies in comparison to the individual classification results. In concept this can be justified by the fact that each of the proposed solutions addresses human activity classification from a different perspective, namely; the CNN Titles of Section III-A addresses it from a spatial perspective, the CNN FVs of Section III-B addresses it from a temporal

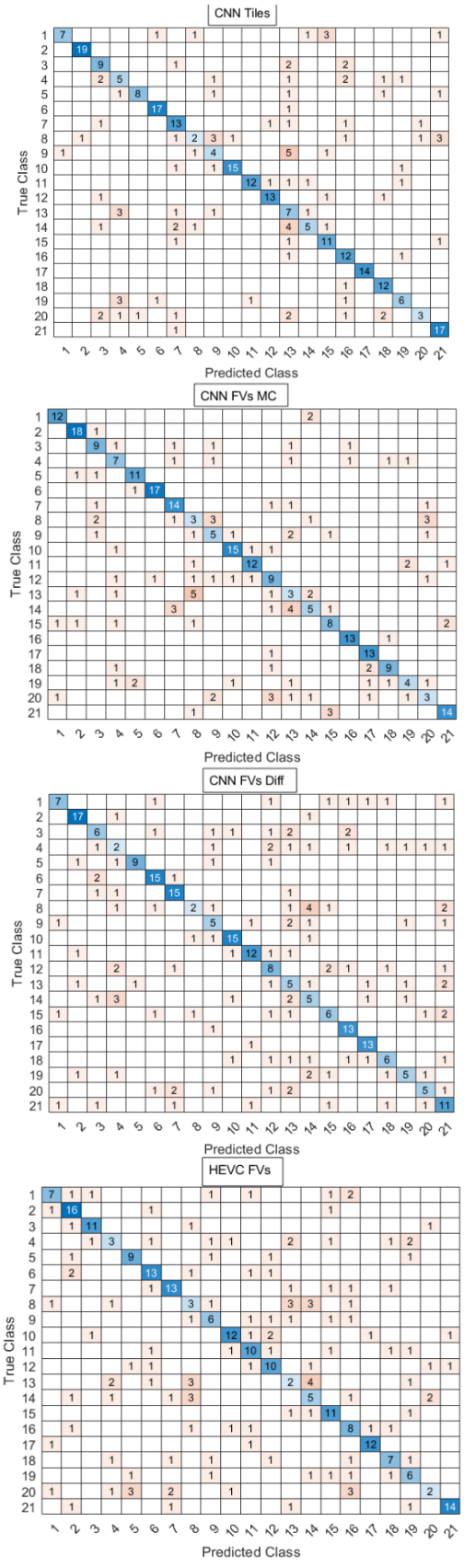

FIGURE 9. Example confusion matrices of the jHMDB dataset using the proposed solutions.

perspective and the HEVC FVs of Section 3.3 addresses it from a spatio-temporal perspective. Combining all of these solutions using score level fusion resulted in the reported classification accuracies.

This observation can be further analyzed by terms of examining the confusion matrices of the individual proposed solutions. Example confusion matrices of the jHMDB dataset are presented in Figure 9. As seen in the matrices, the proposed solutions make correct/incorrect classifications at different human activity classes, thus fusing these solutions resulted in refined classification accuracies in comparison to individual solutions as shown in Figure 10.

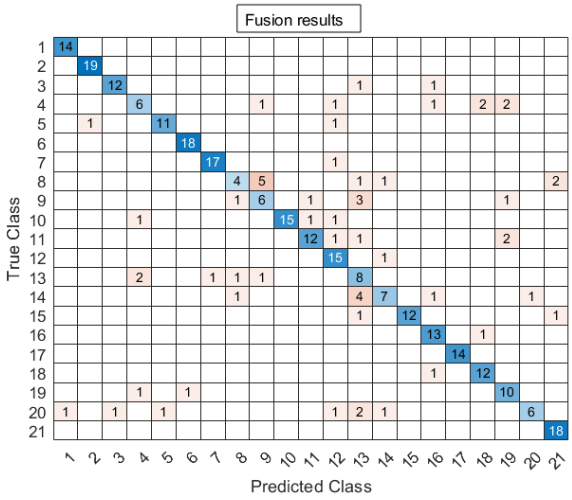

FIGURE 10. Confusion matrix of the jHMDB dataset after score fusion of the results presented in Fig. 7.

As for comparison with existing work, we present the average classification accuracies of the proposed solution against a variety of well-known solutions as summarized by [20].

Please note that the datasets used in this work have been used in human activity recognition for more than a decade and therefore, proposing solutions to further enhance the classification accuracy is becoming a very challenging task. As such, the expectation for improvements in accuracy should be reasonable.

The observations drawn from the results in Tables 3, 4 and 5 are as follows. First, it is clear that the most challenging dataset to classify is HMDB51 followed by jHMDB followed by UCF11. Second, the classification results of the proposed solution are within a reasonable margin of the top results reported in the tables, which gives an indication that the proposed solutions are implemented correctly. Third, the average classification accuracies of the proposed solutions are higher than the reported existing solutions. As mentioned previously, in concept this can be justified by the fact that each of the proposed solutions addresses human activity classification from a different perspective, namely; spatial, temporal and spatio-temporal perspectives. Additionally, the proposed solution contain novel contributions drawn from the video coding domain including HEVC feature variables and

TABLE 3. Comparison of classification accuracy of the proposed solution against existing solutions using the jHMDB dataset.

Method	Accuracy
PoTion[37]	57.0%
P-CNN[38]	61.1%
Action Tubes[39]	62.5%
EleAtt-GRU[40]	62.9%
PA3D[41]	69.5%
MR Two-Stream R-CNN[42]	71.1%
DR2N[43]	71.8%
Generalized Rank Pooling with the improved Trajectory Features[44]	73.7%
Chained MultiStream[45]	76.1%
Temporal Pyramid with the improved Trajectory Features [20]	77.4%
Proposed solution (ViCoMoCoDeL)	**78.5%**

TABLE 4. Comparison of average classification accuracy of the proposed solutions against existing solutions using the HMDB51 dataset.

Method	Accuracy
PoTion [37]	43.7%
C3D [46]	51.6%
PA3D [41]	55.3
Two Stream [47]	59.4%
iTF [48]	61.7%
TDD [49]	63.2%
Temporal Pyramid with the improved Trajectory Features [20]	69.5%
IF-TTN [50]	70%
ARTNet [51]	70.9%
Proposed solution (ViCo-MoCo-DL)	**71.4%**

TABLE 5. Comparison of average classification accuracy of the proposed solutions against existing solutions using the UCF11 dataset.

Method	Accuracy
DEEPEYE [52]	86.1%
Cho et al. [53]	88.0%
Chirag et al. [54]	89.43%
Nazir et al. [55]	93.9%
Ali et al. [56]	95.7%
Bag of Expression (BoE) [57]	96.7%
Temporal Pyramid with the improved Trajectory Features [20]	96.83%
Proposed solution (ViCo-MoCo-DL)	**97.0%**

motion compensation. These novel approaches helped boost the classification accuracies of human motion activities.

VII. CONCLUSION

This work proposed the use of video coding techniques for feature extraction including motion compensation and feature variables from the HEVC coding process. Three solutions were proposed and then fused at a classification score level. In one solution, motion is captured using motion vectors and motion compensation, consequently 4 RGB images are constructed using the underlying motion information. The 4 images are tiled to generate one big image that is inputted to a CNN network for training and classification. In the second solution, each of the constructed 4 RGB images that capture the motion go through a pre-trained CNN to generate feature vectors. These feature vectors are then stacked to generate a feature matrix and used with an LSTM classifier. The third solution uses HEVC to generate feature variables and then use statistical summaries to generate feature vectors which are stacked to generate a feature matrix and used with an LSTM classifier as well.

It was noticed that in comparison to traditional motion capture techniques used in the literature, the proposed use of motion compensation resulted in more precise accumulated image differences and higher classification accuracies.

Using confusion matrices, it was also noticed that the three proposed solutions generate correct/incorrect classification results at different class locations, thus fusing the results using a classification score level worked well. Additionally, the three proposed solutions mainly tackle the spatial, temporal and spatio-temporal aspects of the input videos.

Although the datasets used in this work have been used in human activity recognition for more than a decade and excellent results have been reported in the literature, nonetheless, the use of the proposed solutions reasonably increased the accuracy of the classification on three well-known datasets.

ACKNOWLEDGMENT

This paper represents the opinions of the authors and does not mean to represent the position or opinions of the American University of Sharjah.

REFERENCES

[1] I. A. Lawal and S. Bano, "Deep human activity recognition with localisation of wearable sensors," *IEEE Access*, vol. 8, pp. 155060–155070, 2020.

[2] F. Rustam, A. A. Reshi, I. Ashraf, A. Mehmood, S. Ullah, D. M. Khan, and G. S. Choi, "Sensor-based human activity recognition using deep stacked multilayered perceptron model," *IEEE Access*, vol. 8, pp. 218898–218910, 2020.

[3] D. Thakur, S. Biswas, E. S. L. Ho, and S. Chattopadhyay, "ConvAE-LSTM: Convolutional autoencoder long short-term memory network for smartphone-based human activity recognition," *IEEE Access*, vol. 10, pp. 4137–4156, 2022.

[4] W. Qi, H. Su, and A. Aliverti, "A smartphone-based adaptive recognition and real-time monitoring system for human activities," *IEEE Trans. Hum.-Mach. Syst.*, vol. 50, no. 5, pp. 414–423, Oct. 2020.

[5] I. Alrashdi, M. H. Siddiqi, Y. Alhwaiti, M. Alruwaili, and M. Azad, "Maximum entropy Markov model for human activity recognition using depth camera," *IEEE Access*, vol. 9, pp. 160635–160645, 2021.

[6] H. Ramirez, S. A. Velastin, I. Meza, E. Fabregas, D. Makris, and G. Farias, "Fall detection and activity recognition using human skeleton features," *IEEE Access*, vol. 9, pp. 33532–33542, 2021.

[7] A.-C. Popescu, I. Mocanu, and B. Cramariuc, "Fusion mechanisms for human activity recognition using automated machine learning," *IEEE Access*, vol. 8, pp. 143996–144014, 2020.

[8] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, and X. Lv, "Spatial–temporal pooling for action recognition in videos," *Neurocomputing*, vol. 451, pp. 265–278, Sep. 2021.

[9] B. Sheng, J. Li, F. Xiao, and W. Yang, "Multilayer deep features with multiple kernel learning for action recognition," *Neurocomputing*, vol. 399, pp. 65–74, Jul. 2020.
[10] Y. Ming, F. Feng, C. Li, and J.-H. Xue, "3D-TDC: A 3D temporal dilation convolution framework for video action recognition," *Neurocomputing*, vol. 450, pp. 362–371, Aug. 2021.
[11] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3034–3042, doi: 10.1109/CVPR.2016.331.
[12] U. Ahsan, R. Madhok, and I. Essa, "Video jigsaw: Unsupervised learning of spatiotemporal context for video action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 179–189, doi: 10.1109/WACV.2019.00025.
[13] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 12038–12047, doi: 10.1109/CVPR.2019.01232.
[14] C. Li, Q. Zhong, D. Xie, and S. Pu, "Collaborative spatiotemporal feature learning for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7864–7873, doi: 10.1109/CVPR.2019.00806.
[15] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib, J. A. Khan, and A. A. Abbasi, "Human action recognition using fusion of multiview and deep features: An application to video surveillance," *Multimedia Tools Appl.*, pp. 1–27, Mar. 2020, doi: 10.1007/s11042-020-08806-9.
[16] J.-H. Kim and C. S. Won, "Action recognition in videos using pre-trained 2D convolutional neural networks," *IEEE Access*, vol. 8, pp. 60179–60188, 2020, doi: 10.1109/ACCESS.2020.2983427.
[17] J. C. Stroud, D. A. Ross, C. Sun, J. Deng, and R. Sukthankar, "D3D: Distilled 3D networks for video action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Snowmass Village, CO, USA, Mar. 2020, pp. 614–623, doi: 10.1109/WACV45572.2020.9093274.
[18] K. Muhammad, Mustaqeem, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino, and V. H. C. de Albuquerque, "Human action recognition using attention based LSTM network with dilated CNN features," *Future Gener. Comput. Syst.*, vol. 125, pp. 820–830, Dec. 2021, doi: 10.1016/j.future.2021.06.045.
[19] J. Xu, R. Song, H. Wei, J. Guo, Y. Zhou, and X. Huang, "A fast human action recognition network based on spatio-temporal features," *Neurocomputing*, vol. 441, pp. 350–358, Jun. 2021, doi: 10.1016/j.neucom.2020.04.150.
[20] A. Javidani and A. Mahmoudi-Aznaveh, "Learning representative temporal features for action recognition," *Multimedia Tools Appl.*, vol. 81, no. 3, pp. 3145–3163, Jan. 2022, doi: 10.1007/s11042-021-11022-8.
[21] K. P. Sanal Kumar and R. Bhavani, "Human activity recognition in egocentric video using HOG, GiST and color features," *Multimedia Tools Appl.*, vol. 79, nos. 5–6, pp. 3543–3559, Feb. 2020, doi: 10.1007/s11042-018-6034-1.
[22] A. Kushwaha, A. Khare, and P. Srivastava, "On integration of multiple features for human activity recognition in video sequences," *Multimedia Tools Appl.*, vol. 80, nos. 21–23, pp. 32511–32538, Sep. 2021, doi: 10.1007/s11042-021-11207-1.
[23] K. Singh, C. Dhiman, D. K. Vishwakarma, H. Makhija, and G. S. Walia, "A sparse coded composite descriptor for human activity recognition," *Exp. Syst.*, vol. 39, no. 1, pp. 1–19, Jan. 2022, doi: 10.1111/exsy.12805.
[24] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012, doi: 10.1109/TCSVT.2012.2221191.
[25] O. Issa and T. Shanableh, "CNN and HEVC video coding features for static video summarization," *IEEE Access*, vol. 10, pp. 72080–72091, 2022, doi: 10.1109/ACCESS.2022.3188638.
[26] A. Eltayeb and T. Shanableh, "Data embedding in scrambled video by rotating motion vectors," *Multimedia Tools Appl.*, vol. 81, pp. 25473–25496, Mar. 2022.
[27] T. Shanableh, "HEVC video encryption with high capacity message embedding by altering picture reference indices and motion vectors," *IEEE Access*, vol. 10, pp. 22320–22329, 2022, doi: 10.1109/ACCESS.2022.3152548.
[28] S. Youssef and T. Shanableh, "Detecting double and triple compression in HEVC videos using the same bit rate," *Social Netw. Comput. Sci.*, vol. 2, no. 5, Aug. 2021, doi: 10.1007/s42979-021-00800-8.
[29] T. Shanableh, "Feature extraction and machine learning solutions for detecting motion vector data embedding in HEVC videos," *Multimedia Tools Appl.*, vol. 80, pp. 27047–27066, Sep. 2020, doi: 10.1007/s11042-020-09826-1.
[30] T. Shanableh and M. Hassan, "Predicting split decisions in MPEG-2 to HEVC video transcoding," *Social Netw. Appl. Sci.*, vol. 2, no. 6, Jun. 2020, doi: 10.1007/s42452-020-2909-7.
[31] R. Mahmoud, T. Shanableh, I. P. Bodala, N. V. Thakor, and H. Al-Nashash, "Novel classification system for classifying cognitive workload levels under vague visual stimulation," *IEEE Sensors J.*, vol. 17, no. 21, pp. 7019–7028, Nov. 2017, doi: 10.1109/JSEN.2017.2727539.
[32] M. Hassan, K. Assaleh, and T. Shanableh, "Multiple proposals for continuous Arabic sign language recognition," *Sens. Imag.*, vol. 20, p. 4, Jan. 2019, doi: 10.1007/s11220-019-0225-3.
[33] V. Sze, M. Budagavi, and G. Sullivan, *High Efficiency Video Coding (HEVC), Algorithms and Architectures*. Cham, Switzerland: Springer, 2014.
[34] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2556–2563.
[35] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3192–3199.
[36] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,'" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1996–2003.
[37] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose motion representation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7024–7033.
[38] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN features for action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3218–3226.
[39] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 759–768.
[40] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "Adding attentiveness to the neurons in recurrent neural networks," in *Computer Vision—ECCV 2018* (Lecture Notes in Computer Science), vol. 11213. Cham, Switzerland: Springer, 2018, pp. 135–151.
[41] A. Yan, Y. Wang, Z. Li, and Y. Qiao, "PA3D: Pose-action 3D machine for video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7914–7923.
[42] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Computer Vision—ECCV 2018* (Lecture Notes in Computer Science), vol. 9908. Cham, Switzerland: Springer, 2016, pp. 744–759.
[43] C. Sun, A. Shrivastava, C. Vondrick, R. Sukthankar, K. Murphy, and C. Schmid, "Relational action forecasting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 273–283.
[44] A. Cherian, B. Fernando, M. Harandi, and S. Gould, "Generalized rank pooling for activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1581–1590.
[45] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2923–2932.
[46] D. Tran, L. Bourdev, and R. Fergus, "C3D: Generic features for video analysis," *CoRR*, vol. 2, no. 7, p. 8, 2014.
[47] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.
[48] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 3551–3558.
[49] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4305–4314.
[50] K. Yang, P. Qiao, D. Li, and Y. Dou, "IF-TTN: Information fused temporal transformation network for video action recognition," 2019, *arXiv:1902.09928*.

[51] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1430–1439.
[52] Y. Cheng, G. Li, and H.-B. Chen, "DEEPEYE: A compact and accurate video comprehension at terminal devices compressed with quantization and tensorization," 2018, *arXiv:1805.07935*.
[53] J. Cho, M. Lee, H. J. Chang, and S. Oh, "Robust action recognition using local motion and group sparsity," *Pattern Recognit.*, vol. 47, no. 5, pp. 1813–1825, May 2014.
[54] C. I. Patel, S. Garg, T. Zaveri, A. Banerjee, and R. Patel, "Human action recognition using fusion of features for unconstrained video sequences," *Comput. Electr. Eng.*, vol. 70, pp. 284–301, Aug. 2018.
[55] S. Nazir, M. Yousaf, and S. Velastin, "Feature similarity and frequency-based weighted visual words codebook learning scheme for human action recognition," in *Proc. Pacific-Rim Symp. Image Video Technol.* Cham, Switzerland: Springer, 2017, pp. 326–336.
[56] A. Javidani and A. Mahmoudi-Aznaveh, "A unified method for first and third person action recognition," in *Proc. Electr. Eng. (ICEE), Iranian Conf.*, May 2018, pp. 1629–1633.
[57] S. Nazir, M. H. Yousaf, J.-C. Nebel, and S. A. Velastin, "A bag of expression framework for improved human action recognition," *Pattern Recognit. Lett.*, vol. 103, pp. 39–45, Feb. 2018.

TAMER SHANABLEH (Senior Member, IEEE) was born in Scotland, U.K. He received the M.Sc. degree in software engineering and the Ph.D. degree in electronic systems engineering from the University of Essex, in 1998 and 2002, respectively.

Then, he was a Senior Research Officer with the University of Essex for three years, where he was collaborated with BTexact on inventing video transcoders and then, he joined the Motorola U.K. Research Laboratories and contributed to establishing a new profile within the ISO/IEC MPEG-4 known as the error resilient simple scalable profile. He joined the American University of Sharjah, in 2002. He is currently a professional engineer. He is also a professor in computer science. During the Summer breaks, he was a Visiting Professor with the Motorola Laboratories in five different years. He spent his sabbatical leave as a Visiting Academic with the Multimedia and Computer Vision and Laboratory, Queen Mary, University of London, U.K. He has six patents and authored more than 80 publications, including ten IEEE TRANSACTIONS papers. His research interests include digital video coding and processing and pattern recognition.