3D Human Action Recognition using Hu Moment Invariants and Euclidean Distance Classifier

Fadwa Al-Azzo
System Engineering Department
University of Arkansas at Little Rock
Arkansas, USA

Arwa Mohammed Taqi
System Engineering Department
University of Arkansas at Little Rock
Arkansas, USA

Mariofanna Milanova
Computer Science Department
University of Arkansas at Little Rock
Arkansas, USA

***Abstract*—This paper presents a new model of scale, rotation, and translations invariant interest point descriptor for human actions recognition. The descriptor, HMIV (Hu Moment Invariants on Videos) is used for solving surveillance camera recording problems under different conditions of side, position, direction and illumination. The proposed approach deals with raw input human action video sequences. Seven Hu moments are computed for extracting human action features and for storing them in a 1D vector which is constringed as one mean value for all the frames' moments. The moments are invariant to scale, translation, or rotation, which is the robustness point of Hu moments algorithm. The experiments are evaluated using two different datasets; KTH and UCF101. The classification process is executed by calculating the Euclidean distance between the training and testing datasets. Human action with minimum distance will be selected as the winner matching action. The maximum classification accuracy in this work is 93.4% for KTH dataset and 92.11% for UCF101.**

Keywords*—human action recognition; Hu moment invariants; surveillance camera; Euclidean distance

I. INTRODUCTION

Human motion analysis is an important field of research in computer vision with many applications including surveillance footage, scene realization, user-interfaces, automatic activity recognition and augmented reality. Over the past years, human action recognition in videos has been popularized to have many real-world applications [1]. Thus, human action recognition has found applications across different scientific fields including information technology, artificial intelligence, image processing, acoustics classification, communication, computer diagnosis, and data mining [2].

However, the assortment of realistic video data has given rise to different challenges for action recognition. It has been a challenging problem in the computer vision analysis. The shape descriptors of moment invariants are important in computer vision. There are two types of shape descriptors: contour-based shape descriptors and region-based shape descriptors. Regular moment invariants are part of the most popular and are highly classified as contour-based shape descriptors [3], [4].

Generally, the human action recognition process includes two steps: feature extraction, and classification process. In this paper, we focus on recognizing different human action from video clips of KTH and UCF101 datasets including various environment backgrounds (outdoor, indoor, different views with difference clothes, gender). The video clips are recorded using surveillance camera that is stable with changes of recording conditions like side, position, direction and illumination. These conditions cause the problem of distinguishing the action accurately. To overcome this problem and to improve the recognition accuracy, Hu moments approach introduced by Hu [3] has been used in this work, where the values are invariant with respect to the scale, translation, and rotation. Moment invariants were chosen because they are one of the most important and most used methods in the object recognition field. The shape descriptors of Hu moments feature have been continuously developed and are a powerful tool for image recognition applications.

Seven-moment invariants have been calculated for each frame of KTH and UCF101 video clips, and they are stored as a 1D vector. To constringe our results, we compute the average value of these moments for each frame, and the final average of all these sub-averages has been calculated, which represents the dominant feature of that video clip. These calculations are repeated for the testing dataset also. Euclidean distance is the appropriate method for the classification process to measure the minimum distance between the final average of both the training and testing datasets for each class. Minimum distance indicates the closer human action of testing dataset to human action class of the training data.

The organization of this paper is as follows: *Section IV, A*, a database of different human actions is formed from each dataset. *Section VI, A*, the HMIV is trained based on that database by applying Hu moment invariants algorithm designed for feature extraction. *Section VI, B* a test dataset is applied on the proposed system utilizing the Euclidean distance classifier EDC to recognize the human action. *Section VI, C*, a comparison between the proposed HMIV and spatial-temporal SURF (ST-SURF) technique has been taken into consideration.

II. LITERATURE REVIEW

Human action recognition in video is a significant and challenging problem in computer vision and machine learning. Some of researchers focus on developing the recognition accuracy as in [5] and [6] by using huge and complex benchmark datasets as presented in [7]-[11]. Many different algorithms are used to achieve the best accuracy in the recognition process.

An adaptive multiple kernel learning approach applied in [12] to get the minimum mismatch between distributions from YouTube and consumer videos. While in [13], they used saliency thresholding concept to remove features from non-salient regions, and the remaining features contribute equally to the final representation measures. On the other hand, the moment technique itself implemented successfully in object detection presented in [4] and [14], also in trademark identification existed in [15] as pattern recognition. In general, the basic contribution of researches consisted of using Geometric Invariant Moment (GIM) to recognize objects of captured images.

Many specialized algorithms have been advanced for human action recognition. Computer vision applications include the task of detecting harmonization between two images with the similar scene or entity. [16]. For action recognition in video, both patio and temporal features are needed to represent the actions, while only spatial features, such as SIFT and SURF are needed for object scene recognition on a still image [17]. In action recognition's state-of-the-art, the BoW models are widely used since they have shown the effectiveness of local appearance based descriptors [18], [19].

In comparison with other approaches, Bag of visual word selection is still in its infancy. To extract video descriptors, many researchers have been investigating in tracking major parts of human bodies then extracting features from these regions [20]. However, they needed to setup many hypotheses. These considerations and hypothesis are often demanding. However, methods based on spatiotemporal features are promising for action recognition. Some of them were based on the extraction of low-level optical flows from cuboids [21]. This method gave good results in terms of feature selection and good classifications accuracy [21]. Besides, in [22] a spatiotemporal descriptor called ST-SURF was presented. That work was based on a combination of the speed up robust feature and the optical flow. In [23], the authors presented an algorithm for human action recognition from videos. His method was based on a combination of two feature types extracted from Aligned Motion Images (AMIs). AMIs is a technique for capturing the motion of all video's frames in one image. In addition, [24] was also based on aligned motion images (AMIs), but using three different sorts of AMIs, aligned motion history image (AMHI), aligned motion energy image (AMEI), and aligned gait energy image (AGEI).

It is worth to mention that the proposed work in this paper has been compared with the state-of-art [22], in terms of feature extraction, classification technique, and evaluated accuracy results.

III. HU MOMENTS THEORY

The moment invariants were first introduced by Hu [3]. Hu moments algorithm is chosen to extract image features since the generated features are rotation scale translation. Geometric Moment GM was successfully applied in aircraft identification, texture classification and radar images for optical images matching [25].

Basic terms in the construction of the invariant moments have two steps. First, consider an image that has a gray function $f(x,y)$ having a bounded support and a finite nonzero integral. Second, geometric moment m_{pq} of the digital sampled $M \times M$ image $[f(x,y)]$, which can be computed using (1) [4].

$$m_{pq} = \sum_{x0}^{x=M-1} \sum_{y=0}^{y=M-1} (x)^p \cdot (y)^q \; f(x,y), \tag{1}$$

p,q = 0,1,2,3,….., where p,q are non-negative integers and (p+q) is called the order of the moment.

The moments of f (x, y) are translated by an amount (a, b), which is calculated by (2).

$$\mu_{pq} = \sum_x \sum_y (x+a)^p \cdot (y+b)^q \; f(x,y). \tag{2}$$

Consequently, the central moment μpq can be calculated from (2) by replacing a = - $\bar{x}$, and b = - $\bar{y}$ as

$$\bar{x} = \frac{m_{1,0}}{m_{0,0}}, \bar{y} = \frac{m_{0,1}}{m_{0,0}},$$

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p \cdot (y-\bar{y})^q \; f(x,y), \tag{3}$$

The central moment of the image is invariant to translation, while the scaling invariance can be achieved by normalizing the moments of the scaled image by the scaled energy of the original image that can be computed as stated below.

$$\eta_{pq} = \frac{\mu_{pq}}{\mu^{\gamma}{}_{00}}, \; \gamma = \frac{p+q}{2} + 1,$$

where γ is the normalization factor.

In fact, Hu defined seven values, calculated by normalizing central moments completed order three that are invariant to object scale, position, and orientation. In terms of the central moments, the seven moments are given as shown in (4) [26].

$$\begin{aligned}
M_1 &= \eta_{20} + \eta_{02}, \\
M_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta^2{}_{11}, \\
M_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2, \\
M_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2, \\
M_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + \\
&\quad (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})\left[(\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2\right], \\
M_6 &= (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] + \\
&\quad 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}), \\
M_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})\right] + \\
&\quad (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})\left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} - \eta_{03})^2\right].
\end{aligned} \tag{4}$$

IV. PROPOSED TECHNIQUE

The proposed HMIV approach aims to detect a specific human action from videos of *N* frames. It includes extracting the features of the training and testing datasets using HMI algorithm. The extracted features of each action for both training and testing datasets are constringed to one magnitude, representing the distinctive features of that action. The classification process is the next step which depends on the Euclidian distance classifier between the training and testing datasets. Eventually, the human action of the minimum distance value would be selected as a matching action. Fig. 1 shows the block diagram of HMIV approach.

A. Data Acquisition

Data have been acquisition using KTH and UCF101 datasets. For each action, different recording conditions are covered, indoor, outdoor, various outfits, and gender. The input data is represented as frames sequences. In other words, the input is considered as 3D characteristics; (Height) × (Width) × (Number of frames). Fig. 2 shows the two used datasets.

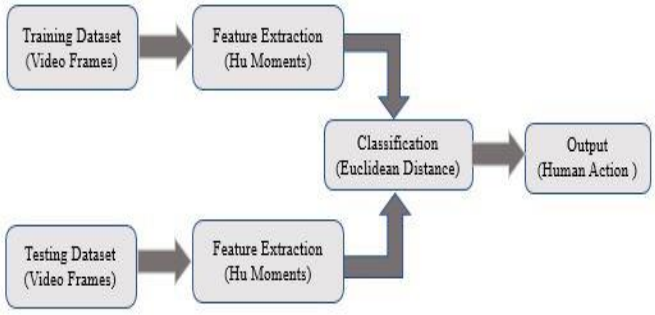


Fig. 1. Block diagram of the proposed HMIV approach

a. KTH dataset

b.UCF 101 dataset

Fig. 2. Two datasets including different human actions under various conditions

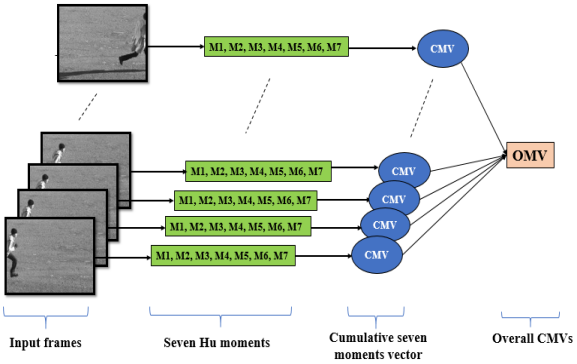


Fig. 3. Structure of feature extraction process

B. Feature Extraction

Features extraction process is a method of image transformations, which is capable to transfer high-dimension feature to the low-dimension feature vector. In another word, the feature extraction accomplishes dimensional compression, while preserving the essential information, which is most characteristic and necessary to the image [27]. Features selection and extraction are an outstanding process amongst the most significant strides in image recognition, which could impact on coming recognition process stages [28].

Indeed, feature extraction process includes computing seven Hu moments for each frame, and all moment's values are concatenated into a 1D vector.

Under those circumstances, we calculated a vector of seven Hu moments for each frame. To address this issue, the average of those moments are computed as cumulative moments value (CMV) as shown in (5). To be able to accomplish accurate and fast calculation in the proposed algorithm, an average of all cumulative Hu moments values (CMVs) are constringed as overall moments value (OMV) for all the target frames carry out by (6). In fact, OMV grantees the dominant features that extracted from all the input frames for a specific human action. Fig.3 illustrates features extraction process.

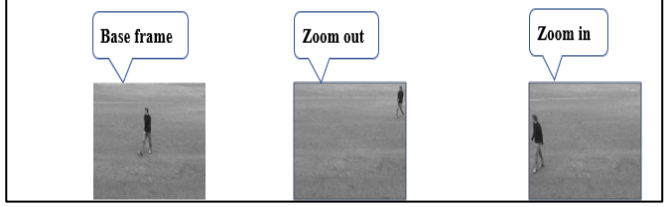
$$\mu_{i,j} = \frac{\sum_{k=1}^{K} M_{i,j,k}}{K}, \qquad \forall i, \forall j, \tag{5}$$

where $\mu_{i,j}$ denotes the cumulative moments values, i is the number of human action's classes, and j represents numbers of frames.

$$\omega_i = \frac{\sum_{j=1}^{J} \mu_{i,j}}{J}, \qquad \forall i, \tag{6}$$

where ω_i symbolizes the overall moments value for each action, while J is the number of frames.

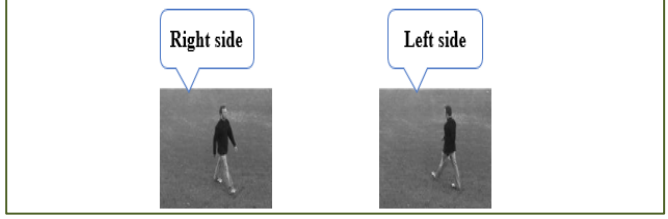
	M1	M2	M3	M4	M5	M6	M7
Base frame	2.3062	4.8459	7.2154	6.8325	14.0558	9.2554	-14.4625
Zoom out	2.3057	4.8389	7.2085	6.8319	14.0518	9.2513	-14.4934
Zoom in	2.3163	4.8638	7.2393	6.8607	14.1057	9.2927	-14.4268



(a)

(b)

	M1	M2	M3	M4	M5	M6	M7
Right side	2.2118	4.6647	6.9392	6.5435	13.4806	8.8759	-13.8491
Left side	2.2216	4.6852	6.9704	6.5718	13.5377	8.9144	-13.9061



(c)

Fig. 4. The power of Hu moment invariants identification

C. Hu moments analysis on human action

In human actions situation, it is essential to deal with an effective and valuable concepts in motion representation are based on HMI. For an instant, video sequences of walking action are taken, obviously, the HMI have such stationary values with a minor disparity in digits. As shown in Fig. 4 (a), it is back clearly that any individual moment of the base frame (M1 or M2 or … M7) almost preserves the same value in the zoom out and zoom in camera shot, even though the tracking person has various scaling. The zoom in and zoom out frames confirm the robustness of HMI; their moments are nearly identical (Fig. 4 (b)). Besides, the different positions (right or left) almost have no fluctuations on moments values individually (Fig. 4 (c)). For example, M1 in right side frame is 2.2118, which is just near to M1 value of the left side frame 2.2216). That highly demonstrates the efficiency of HMI theory on preserving the extracted features. In addition, person rotation could alter the image function more or less. Nonetheless, the moment invariants save varying while the person is rotated. Indeed, the theory strength relies on (4) entities, which cover all the possible recording conditions from the surveillance camera.

D. Classification Process

The distance measurement or similarity between images is an essential and an open issue in the computer vision and machine learning. The most commonly used distance is Euclidean distance, which converts images into vectors according to the gray levels of each pixel, and then compares intensity differences pixel by pixel. Deriving the Euclidean distance between two data points involves computing the square root for the sum of the differences squares between corresponding values, as described in [29].

Many applications in machine learning have commonly used the Euclidean distance, for an instance, K-Nearest Neighbor, K-Means Clustering, and the Gaussian kernel. Calculating the Euclidean distance can be significantly enhanced by taking benefit of the distinct instructions for performance matrix multiplications [30].

The Euclidean distance can be written in terms of a matrix multiplication that requires some reworking of the distance equation. In this work, recognition of human actions basically depends on Euclidean distance concept, which is easy to apply and less computational complexity. The classification process is performed using a convenient Euclidean distance classifier (EDC).

A similarity metric is based on the EDC, which computes the differences between OMV of the testing dataset and OMVs for different action classes of the training dataset, as adopted in (7):

$$D_{optimum} = \min\left(\sqrt{\left((\omega)_{Testing} - (\omega_i)_{Training}\right)^2}\right), \qquad (7)$$

where $D_{optimum}$ is the minimum distance between the testing and training actions.

Fig. 5 demonstrates the proposed classification strategy. As described in *Section VI. B*, OMV for each action of the training dataset have been computed. Subsequently, OMV for the testing dataset has been calculated. The classifier EDC uses (7) to find the minimum difference value between the testing and training OMVs for each action. The action with the minimum value $D_{optimum}$ would be recognized as the closed action. Actually, $D_{optimum}$ refers to a severe convergence between the training action features and the testing one.

V. SURF VS. HMI

SURF algorithm (Speeded Up Robust Features) has been presented in [16]. It is needed for object scene recognition on a still image, and for extracting spatiotemporal features from videos. SURF can select a set of features from a dataset [17].

These features are tested to examine their ability for classifying a human action.

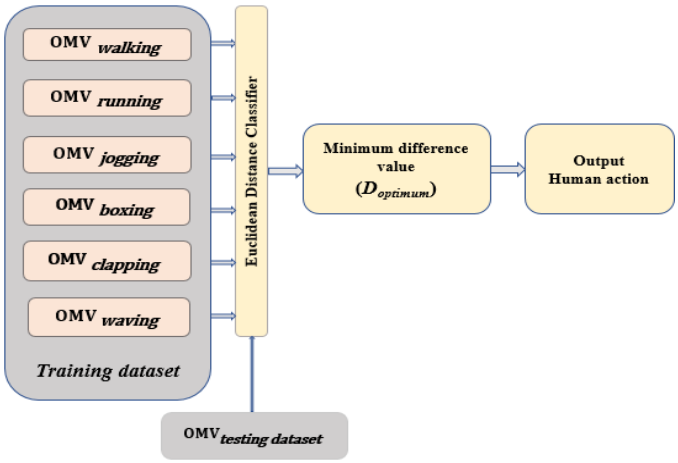


Fig. 5. Block diagram of classification process

The SURF detector includes Hessian-based detectors, which are more steady and repeatable than their Harris-based counterparts. Also, common calculations, like the different of Gaussian (DoG), provide a high speed at a low cost in terms of lost accuracy. In contrast, SURF descriptor defines a spreading of Haar wavelet responses inside the interest point neighborhood [16].

As a part of this research, an explicit comparison between SURF algorithm presented in [22] and our proposed HMIV has been introduced. The work in [22] goaled for detecting human actions based on SURF method. The procedures were summarized in three steps. First, video sequences were segmented in frame packets and a group of interest points. Second, the interest points ST-SURF (Spatial-temporal SURF) were localized and extracted from all training videos. After that, the extracted ST-SURFs were gathered via K-means clustering algorithm. The video clips were characterized as a K-bins histogram of the quantized descriptors "bag of spatiotemporal visual words" BoSTVW. Lastly, an SVM (support vector machine) classifier was trained by means of these histograms. The results of ST-SURF were evaluated using KTH and UCF sports datasets.

On the other hand, the proposed HMIV in this paper aims for human actions recognition; it is built depending on the HMI algorithm for the interest feature extraction, and Euclidean distance classifier EDC for the classification process. This model serves a surveillance camera system that is used to record various videos of human actions under different environments (indoor and outdoor), and different conditions of side, position, direction and illumination. In fact, HMI have an effective efficiency of maintaining their probative values, even though there are various changes in scale, translation, and rotation. Further, EDC classifier is easy to apply, it computes the difference between OMV from each training and testing action. As result, the minimum difference value indicates the winner human action. The datasets in this work are KTH and UCF101.

TABLE I. HU MOMENTS FOR WALKING ACTION FRAMES OF KTH DATASET

Hu moments	M1	M2	M3	M4	M5	M6	M7
Fram1	2.5380	0.3045	0.9058	0.5306	0.4355	0.1829	-15.6406
Fram2	2.5381	0.3046	0.9061	0.5308	0.4359	0.1832	-15.6409
Fram3	2.5381	0.3046	0.9060	0.5308	0.4358	0.1832	-15.6408
Fram4	2.5385	0.3057	0.9079	0.5320	0.4384	0.1849	-15.6427
Fram5	2.5385	0.3060	0.9085	0.5322	0.4392	0.1853	-15.6439
Fram6	2.5384	0.3063	0.9096	0.5322	0.4398	0.1854	-15.6454
Fram7	2.5388	0.3083	0.9146	0.5340	0.4450	0.1883	-15.6513
Fram8	2.5388	0.3087	0.9155	0.5339	0.4453	0.1884	-15.6509
⋮							
Fram48	2.5279	0.2827	0.8679	0.4999	0.3741	0.1413	-15.5974
Fram49	2.5277	0.2816	0.8662	0.4992	0.3726	0.1401	-15.5993
Fram50	2.5271	0.2797	0.8632	0.4975	0.3686	0.1374	-15.5986

VI. EXPERIMENTAL RESULTS

The experiments have been evaluated using two different datasets with gray and color resolutions. The proposed HMIV performance has been estimated under different surveillance camera recording conditions of a side, position, direction, illumination, and an environment. KTH dataset is one of the widely common used datasets, it is presented in [31]. While UCF101 datasets are obtained from [32]. The inputs of the proposed HMIV include six different human actions; walking, running, jogging, boxing, handwaving, and handclapping for KTH dataset. On the other hand, for the UCF 101 dataset, baseball pitch, basketball, bench press, biking, and balance beam are chosen. These actions are considered by way of a database for model training.

A. Training

In this stage, the model was trained using the HMI algorithm. During model training, the feature extraction process was executed for six different categories. Using HMI, many conditions should be considered, such as the various side, positions, illumination including person shadow to guarantee the performance. As mentioned in Section IV. B, at the end of this stage, the most important features were constringed as one dense value, which is OMV for each training action.

By way of example, Table I illustrates the moments for walking action frames of KTH dataset. It shows the effective power of HMI weights with person movement and action displacement. These weights represent the dominant features of frame images. As labeled in Table I, it is obvious that the moments of frame1 (M1, M2,…,M7) nearly keep the equivalent value for the rest frames, excepting minor differences in decimal digits. In addition, Fig. 6 shows an example of the salient HMI features of KTH actions, despite there is a clear convergence of their moment's values.

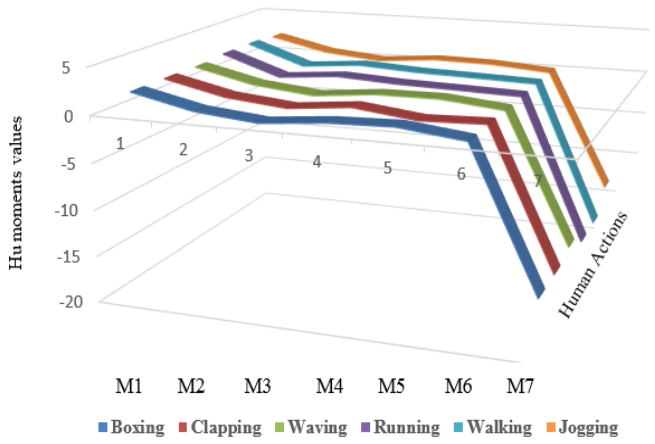


Fig. 6. Hu moment invariants for different human actions of KTH dataset

B. Testing and Results

For the proposed HMIV evaluation, comprehensive investigations have been implemented on KTH and UCF101 datasets. To realize greatest predictable classification accuracy, sets of human actions frames with different conditions and environments are tested. In a like manner of the training section, the testing dataset is processed under HMI algorithm for extracting action features and obtaining OMV. Subsequently, OMV of testing dataset is examined with all OMVs of training dataset classes, to figure out which action is the closest matching one. EDC would be prepared for that classification process by calculating the minimum difference value adopted as (6).

For evaluation purpose, the experimentation results are carried out using KTH and UCF101 datasets. It is verified that the designed HMIV displays promising results. As shown in two Figs. 7 and 8, confusion matrices include the classification accuracy for per human action. For KTH dataset in Fig. 7, jogging action has the maximum classification accuracy about 96%, compared with handwaving action accuracy which is the minimum among them about 88%. Overall, the KTH dataset reaches an aggregate classification accuracy of 93.4%. Moreover, a maximum classification accuracy in UCF101 basketball action reaches to 94.55%, while the minimum accuracy runs to 89.23% at bench press action. The aggregate classification accuracy arrives at 92.11% for the UCF101 dataset.

The classification error rate appears in Fig. 9 (a,b) for KTH and UCF101 respectively. By observing error results, for KTH dataset actions, handwaving has the maximum error value of about 0.12. In contrast, the minimum error has been recorded in jogging around 0.04. But regarding to UCF101 dataset actions, the maximum error value is noted in bench action for about 0.1077, while the minimum one is belonged to basketball action of 0.0564.

C. Comparative Studies

As described in section V, the designed descriptor using ST-SURF was compared with the proposed HMIV. Comparing with the results driven by the best classification accuracy result of the ST-SURF, the HMIV approach achieves 93.4% for KTH dataset better than the 88.2% reported using Spatiotemporal SURF (ST-SURF) presented in [22]. Besides, outperforming the accuracy results related to UCF dataset in HMIV is 92.11%, whereas in ST-SURF was 80.7% as illustrated in Table II.

VII. Conclusion

In this paper, our contribution toward exhibiting HMIV qualification is in the feature extraction stage according to SURF method. A MATLAB code of the *detectSURFFeatures* function is implemented on KTH dataset of walking action. It is important to mention that SURF method cannot recognize the action features when a part of the human body disappears or is hidden from view In contrast, HMIV captures the features of the same case. Fig. 10 confirms this status; SURF descriptor display no feature detect, while our proposed HMIV gives the ordinary seven moments.

The work of proposed HMIV in this paper has been compared with other state-of-the-art techniques. Table III clarifies the classification accuracy results of the human actions literature works in comparison with ours. Overall, the proposed HMIV demonstrates a considerably improved performance compared with other existing state-of-the-art methods.

In this paper, we present and demonstrate the advantage of the HMIV model for human actions recognition of video frames sequence. It gives a basic usefulness of solving surveillance camera recording problems, such as various position (right, left, forth and back), illumination (indoor, outdoor, shadow), and disorganized environment (gender, outfits). Environment changing highly effects on the feature extraction of human actions. It is worth to mention that the Hu moment invariants HMI have such a distinguished power on preserving the extracted features from images sequences. Two different datasets have been used in our approach, KTH, and UCF101 with various human actions. The designed descriptor is based on Hu moment invariants HMI algorithm. The proposed feature extraction process consists on computing seven moments as features that are invariants to scale, translation, and rotation. Then, the moments of each frame are mapped into a 1D vector space for each action class of the training and testing datasets. To reduce the dimension of feature vector into one intensive value, the average of vector's values is calculated as cumulative moments value (CMV).

Because datasets (training and testing) are as 3D input data, and each dataset has multi-actions, so we summarize CMVs into overall moments value OMV. Whereas OMV considers as the dominantly interesting features that extracted from all input frames for a specific human action. Afterward, the recognition process in this work is employed using an appropriate Euclidean distance classifier (EDC).

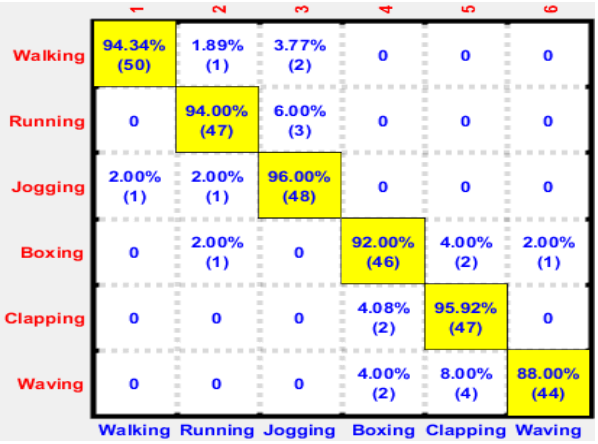


Fig. 7. Confusion matrices for classification accuracy per human action for KTH dataset

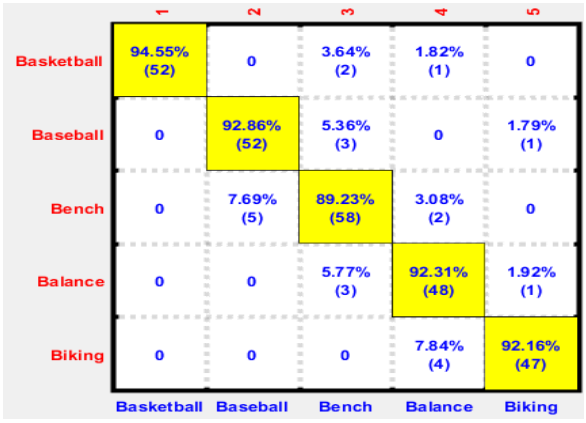


Fig. 8. Confusion matrices for classification accuracy per human action for UCF101 dataset

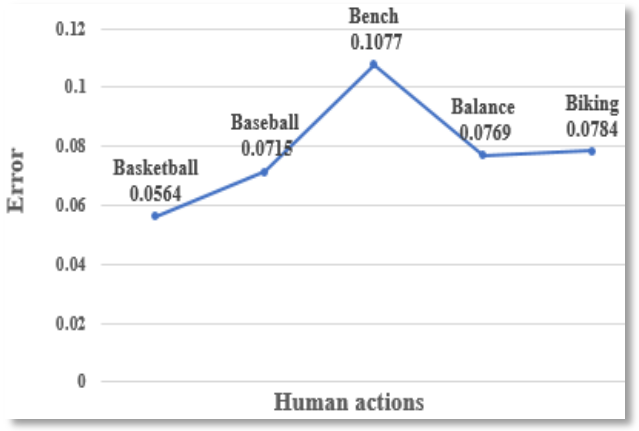
TABLE II. COMPARISON CLASSIFICATION ACCURACY OF PROPOSED HMIV AND ST-SURF

	KTH	UCF
ST-SURF	88.2 %	80.7%
HMIV	**93.4%**	**92.11%**

TABLE III. COMPARISON CLASSIFICATION ACCURACY OF PROPOSED HMIV WITH THE STATE-OF-THE-ARTS

KTH		
Schu¨ldt et al. [31]	Local SVM Approach	71.7%
Doll´ar et al. [33]	Sparse Spatio-Temporal Features	81.2%
Niebles et al. [34]	Unsupervised learning method	83.3%
Jhuang et al. [35]	A biologically-motivated system	91.7%
Ji et al. [36]	3D Convolutional Neural Network	90.2%
HMIV	**Hu Moment Invariants on Video**	**93.4%**
UCF sports		
Wang et al. [37]	Local spatio-temporal features	85.60%
Kovashka et al. [38]	Discriminative Space-Time Neighborhood Features for	87.27%
Arac et al. [39]	Lagrangian Descriptors	89.97%
HMIV	**Hu Moment Invariants on Video**	**92.11%**

(a)



(b)

Fig. 9. Error estimation of the proposed HMIV for: (a) KTH, and (b) UCF101 datasets

SURF	No features						
HMIV	M1	M2	M3	M4	M5	M6	M7
Features	2.3090	4.8449	7.2162	6.8397	14.066	9.2621	-14.517

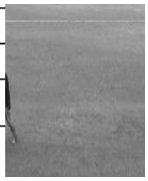


Fig. 10. SURF descriptor shows no feature detect, but HMIV obtains moments features when a part of human body disappears

From the other point of view, we present a comparison between the proposed HMIV and ST-SURF (Spatial-temporal SURF), in terms of the used features extraction technique and classification accuracy. The proposed approach in this paper reaches to 93.4% for KTH dataset, which is better than 88.2% reported via ST-SURF. In addition, HMIV performs more improved than ST-SURF. The accuracy result correlated to HMIV for UCF dataset is 92.11%, however in ST-SURF had 80.7%.

As a future task, we are looking forward to investigating approaches for objects detection, EEG signal classification, and facial expression recognition.

VIII. DISCUSSION

In this paper, human actions are recognized by the implemented HMI algorithm in features extraction and EDC classifier for the recognition process. They are commonly utilized in object recognition due to their discriminations strength and robustness. As has been noted in the confusion matrices, the average accuracy over each of the six actions for KTH and five sports actions of UCF101 datasets is satisfied. HMIV produced good results, in terms of an average classification accuracy of 93.4% for KTH dataset. However, less accuracy is observed in the handwaving action, which is 88%. It is obviously clear that handwaving is similar with handclapping action mostly (8% error ratio), while less similarity with boxing action (4%), because it basically depends on hands motions. In addition, HMIV approach highly discriminates the jogging action with best existing accuracy of 96%. This results due to the fact that the extracted features of this action have a lack of correspondence among other actions, except a small error ration of 2% with running and walking actions.

Besides, for UCF101 dataset has an average recognition accuracy of 92.11%. The bench action has the minimum accuracy of about 89.23%, there is some matching features with the baseball action with 7.69% error ratio. Whereas, basketball is the best recognized action having an accuracy of 94.55%. In other words, HMI algorithm with the EDC could highly capture its features accurately. Lastly but not least, comparing with results driven by the state-of-the-art existing methods.

REFERENCES

[1] H. A. Abdul-Azim and E. E. Hemayed, "Human action recognition using trajectory-based representation," *Egypt. Informatics J.*, vol. 16, no. 2, pp. 187–198, 2015.

[2] J. Han, P. Yang, and L. Zhang, "Object Recognition System of Sonar Image Based on Multiple Invariant Moments and BP Neural Network," vol. 7, no. 5, pp. 287–298, 2014.

[3] M. K. Hu. Visual pattern recognition by moment invariants. IRE Trans. Inform. Theory, IT(8), 1962.

[4] M. Rizon *et al.*, "Object Detection using Geometric Invariant Moment," *Am. J. Appl. Sci.*, vol. 3, no. 6, pp. 1876–1878, 2006.

[5] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In CVPR, 2012.

[6] H. Wang, A. Kl¨aser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In CVPR, 2011.

[7] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos in the wild. In CVPR, 2009.

[8] N. Niebles, C. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In ECCV, 2010.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In CVPR, 2008.

[10] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. MVA, 24(5):971–981, 2012.

[11] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In ICCV, 2011.

[12] L. Duan, D. Xu, I. W. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In CVPR, 2010.

[13] E. Vig, M. Dorr, and D. Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In ECCV, 2012.

[14] Yaakob, S.N., P. Saad and A.H. Abdullah, 2005. Insert recognition using fuzzy ARTMAP. Proc. Intl. Conf. Robotics, Vision, Information and Signal Processing ROVISP2005, pp: 679-684.

[15] Saad, P., 2004. Feature extraction of trademark images using geometric invariant moment and zernike moment- a comparison. Chiang Mai J. Sci., 31: 217-222.

[16] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3951 LNCS, pp. 404–417, 2006.

[17] D. H. Nga and K. Yanai, "A dense SURF and triangulation based spatio-temporal feature for action recognition," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8325 LNCS, no. PART 1, pp. 375–387, 2014.

[18] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, 2008.

[19] H. Riemenschneider, M. Donoser, and H. Bischof, "Bag of Optical Flow Volumes for Image Sequence Recognition," *Procedings Br. Mach. Vis. Conf. 2009*, p. 28.1-28.11, 2009.

[20] M. Mojarrad, M. Dezfouli, and A. Rahmani, "Feature's Extraction of Human Body Composition in Images by Segmentation Method," *World Acad. Sci. Eng. Technol.*, pp. 267–270, 2008.

[21] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR*, 2008.

[22] Z. Yao, Z. Zhang, M. Hu, and Y. Wang, "Advances in Multimedia Information Processing – PCM 2013," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8294, no. 61210006, pp. 79–87, 2013.

[23] M. Milanova and S. Al-ali, "Human action recognition using combined contour-based and silhouette-based features and employing KNN or SVM classifier," no. April 2016, 2015.

[24] S. Al Ali and M. Milanova, "Human action recognition in videos using structure similarity of aligned motion images," *Int. J. Reason. Intell. Syst.*, vol. 6, no. 1/2, p. 71, 2014.

[25] Khotanzad, A., and Y.H. Hong, 1990. Invariant image recognition by zernike moments. IEEE Trans. Pattern Analysis and Machine Intelligence, 12: 489-497.

[26] A. Khotanzad and J. Lu, "Classification of invariant image representation," IEEE Transaction on Acoustics, Speech, and Signal Processing, vol. 38, 1990.

[27] Y. Zilu, L. Jingwen and Z. Youwei, "Facial expression recognition based on two dimensional feature extraction Signal Processing", ICSP 9th International Conference on IEEE, (2008), pp. 1440-1444.

[28] V. Sugumaran, V. Muralidharan and K. I. Ramachandran, "Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing", Mechanical Systems and Signal Processing, vol. 21, no. 2, (2007), pp. 930-942. [

[29] J. Li and B.-L. Lu, "An adaptive image Euclidean distance," Pattern Recognit., vol. 42, no. 3, pp. 349–357, 2009.

[30] C. Mccormick and U. S. Differences, "Fast Euclidean Distance Calculation with Matlab Code," pp. 1–6, 2017.

[31] C. Schuldt, L. Barbara, and S.- Stockholm, "Recognizing Human Actions : A Local SVM Approach ∗ Dept . of Numerical Analysis and Computer Science," Pattern Recognition, 2004. ICPR 2004. Proc. 17th Int. Conf., vol. 3, pp. 32–36, 2004.

[32] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 human actions classes from videos in the wild," CoRR, vol. abs/1212.0, no. November, pp. 1–7, 2012.

[33] E. Florin and S. Baillet, "Behavior Recognition via Sparse Spatio-Temporal Features," Neuroimage, vol. 111, pp. 26–35, 2015.

[34] J. C. Niebles, H. Wang, and L. Fei-Fei "Unsupervised learning of human action categories using spatial temporal words," International Journal of Computer Vision, 79(3):299–318, 2008.

[35] H. Jhuang, T. Serre, L. Wolf, and T. Poggio,"A biologically inspired system for action recognition,"ICCV, pp. 1–8, 2007.

[36] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," Pattern Anal. Mach. Intell. IEEE Trans., vol. 35, no. 1, pp. 221–231, 2013.

[37] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2046–2053, 2010.

[38] C. Mccormick and U. S. Differences, "Fast Euclidean Distance Calculation with Matlab Code," pp. 1–6, 2017.

[39] E. Acar et al, "Action Recognition using Lagrangian Descriptors," IEEE MMSP, pp. 360-365, 2012.