ISSN: 0067-2904

Human Action Recognition Based on Bag-of-Words

Riyadh Sahib Abdul Ameer*, Mohammed Al-Taei
Department of Computer Science, Al-Nahrain University, Baghdad, Iraq

Received: 17/8/ 2019 Accepted: 30/9/2019

Abstract

Human action recognition has gained popularity because of its wide applicability, such as in patient monitoring systems, surveillance systems, and a wide diversity of systems that contain interactions between people and electrical devices, including human computer interfaces. The proposed method includes sequential stages of object segmentation, feature extraction, action detection and then action recognition. Effective results of human actions using different features of unconstrained videos was a challenging task due to camera motion, cluttered background, occlusions, complexity of human movements, and variety of same actions performed by distinct subjects. Thus, the proposed method overcomes such problems by using the fusion of features concept for the development of a powerful human action descriptor. This descriptor is modified to create a visual word vocabulary (or codebook) which yields a Bag-of-Words representation. The True Positive Rate (TPR) and False Positive Rate (FPR) measures gave a true indication about the proposed HAR system. The computed Accuracy (A_r) and the Error (misclassification) Rate (E_r) reveal the effectiveness of the system with the used dataset.

Keywords: Human Action Recognition (HAR), feature extraction, action detection, Bag-of-Words (BoW).

تمييز افعال البشر استناداً على حقيبة الكلمات

رياض صاحب عبد الامير*، محمد صاحب الطائي
قسم علوم الحاسوب ،جامعة النهرين، بغداد ، العراق

الخلاصة

أن التعرف على النشاط البشري قد اكتسب الكثير من الاهتمام بسبب تطبيقاته الواسعة التي شملت: أنظمة المراقبة، متابعة المرضى، وكذلك مجموعة متنوعة من الأنظمة التي تنطوي على تفاعل البشر مع الأجهزة الإلكترونية مثل واجهات الكمبيوتر–البشرية. ان مراحل نمذجة النشاط البشري على الحاسبة تشمل: تجزئة الجسم ، واستخراج المعالم ، واكتشاف الإجراء ، ثم التصنيف أو التمييز. حيث تُعد النمذجة الفعالة للأفعال البشرية باستخدام ميزات مختلفة من مقاطع الفيديو غير المقيدة مهمة صعبة بسبب حركة الكاميرا، الخلفية المزدحمة، وتعقيد الحركات البشرية، وتنوع اشكال نفس الحركة التي يؤديها اشخاص مختلفون. ولذلك ، فأن عملنا المقترح قد استخدام مفهوم دمج الميزات في نموذج واصف قوي التأثير. وهذا الواصف استخدم لإنشاء مفردات الكلمة الواحدة (او ما يسمى دفتر الشفرات) والذي بدوره يقوم بانتاج تمثيل حقيبة الكلمات. ان استخدام مقاييس المعدل الايجابي الحقيقي والمعدل الايجابي الكاذب اعطت مؤشراً حقيقياً حول النظام المقترح، كما وكشفت الدقة المحسوبة ومعدل الخطأ عن فعالية النظام مع مجموعة البيانات المستخدمة.

*Email: riyadhsahib@gmail.com

1. INTRODUCTION

Human action recognition (HAR) is a field of science related to computer vision researches, which is ranged from a simple limb movement to joint complex movement of multiple limbs and human body. This process is a dynamic related to time domain and, thus, it is usually found in a video that is lasting for few seconds [1]. The importance of HAR was coming from its extended relation to various applications that may refer to the subject of personality analysis and identification, leading to predict the next behavior of humans. This makes such subject an important in different fields, especially those related to security including live visual surveillance and video retrieval, etc. An important issue in HAR research is that most studies concentrated on features representation of human action [2].

However, there are a lot of issues related to HAR for videos including:

1. The main issue in HAR is the angle of camera from which to capture videos, as this angle stays in moving. Therefore, the operation of HAR must be stable to the change in the camera's angle movement.
2. Illumination changes and the effect of background occlusions.
3. Different challenges remain at HAR due to the variations between intra-class and inter-class.
4. Changes of human appearance are because of the method of achieving actions that are based on changing the area on which the action is achieved. Also, the clothes show a significant part in human appearance, as well as objects that they hold with. Hence, a case study should be conducted to distinguish the human action related with the human appearance [3].

The bag-of-words (BoW), so-called bag of visual words, is the common feature symbolization method used for document symbolization in the retrieval of information. This method was used by Sivic and Zisserman in the fields of image and video retrieval. Practically, it showed promising efficiency for retrieval tasks and image annotation. The feature extraction procedure in BoW is typically dependent on tokenizing the found-out keypoints to create a visual word vocabulary (or codebook). Thus, the visual word vectors of an image are containing the absence or presence information of each visual word in that image, such as the number of keypoints in the equivalent cluster, which is a visual word. The procedure for extracting features in BoW requires the following points, sequentially [4]:

(a) Automatic detection of points in regions of interest.
(b) Computation of local descriptors for these points.
(c) Formulation of visual vocabulary by quantizing those descriptors into words.
(d) Creating visual features of BoW from the frequencies of the vocabulary configured for each image as histograms.

Figure-1 illustrates the steps to create and extract BoW features, in which the used feature is the orientation of pixel.

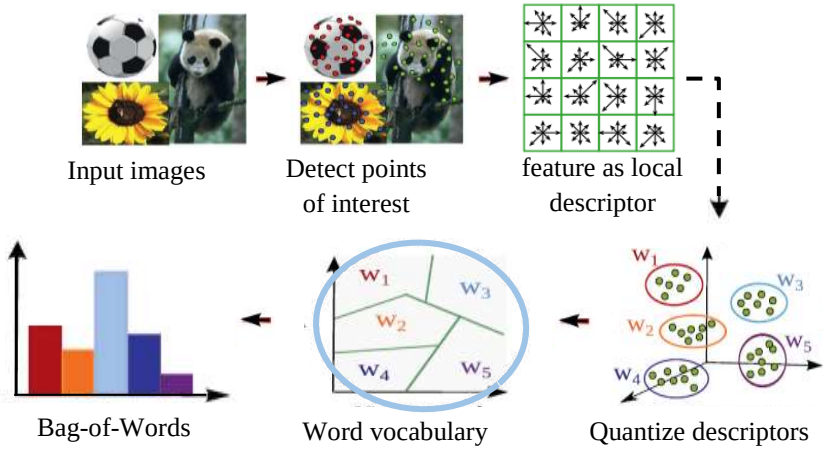


Figure 1-Bag-of-words.

2. RELATED WORK

For the last few decades, HAR has been fully examined. And yet many investigative efforts have been devoted to HAR, but it remains a difficult task because of the appearance differences and subject's movement habits, as well as the variations of viewing angle and changes in illumination.

Reliable segmentation and tracking remain open research problems of incapability to achieve acceptable performance which leads to weak occlusion presences [5].

Masato et al. [6] presented a new method for incohesive motion recognition from a video series. In this method, they used time series spatio temporal intensity gradients within a Space Time patch (ST-patch). Their method was able to distinguish multiple-class motion patterns with a detection rate of about 80%. Also, the detection rule of incoherent motions was 100% with a false positive rate of less than 10%. Kim et al. [7] submitted a technique for distinguishing human actions from a query of a single action video. Their scheme was based on the ordinal measurement of cumulative movement, where this measurement is active at variations of appearances. Song et al. [8] used the BoW representations relied on the quantization of vectors on local spatial-temporal features, because of the good performance and simplicity of these representations. Therefore, a localized, continued and video representation probabilistic was discovered. This makes the representation readily applicable as an input to most discriminative classifiers, such as the nearest neighbor schemes and the kernel methods. Eweiwi et al. [9] introduced a model on temporal key poses for HAR, where Motion History Images (MHI) and Motion Energy Images (MEI) temporal patterns are applied for recognizing human action videos. They combined both methodologies to extract a new representation of temporal key poses. The introduced approach is computationally efficient, robust with respect to parameter selection, and straight forward to implement, as it builds on well-established and understood concepts. Victor et al. [10] presented an approach in videos based on spatiotemporal human object interactions for action recognition, where temporal and spatial evolution of relationship between human and objects is was defined and the position of human and objects are recognized in video frames. The goal is to translate information about relative movement between human and object that changes with time. Then, the features are obtained from identified window of human and object in each frame. Lo et al. [11] extracted local descriptors in a video depending on two concepts, one is using objects motion boundary and the other is the motion boundary trajectories resulting from human action recognition that are obtained from videos. They compared the performance of the proposed motion boundary trajectory approach with many human action benchmark datasets and found that the proposed approach gives improved recognition results. Wang et al. [12] submitted a precise and efficient background subtraction technique. It was concentrated on decreasing the data dimensionality of an image frame based on compressive sensing. Then a sparse representation is applied to create the current background from a set of background images. The proposed method is validated through multiple challenging video sequences. Their results demonstrated that the performance of the approach is comparable to those of the existing classical background subtraction techniques. Gaba et al. [13] introduced a motion detection approach by understanding the importance of identifying moving objects. It has a proficient identification of moving objects, which is adjustable to noise from the background and differences in brightness. This approach is a pixel dependent and non-parameterized, that is dependent on the first frame to form the model. The proposed algorithm was tested on several open source videos by imposing a single set of variables to overcome shortcomings of relevant and recently developed techniques. Sharif et al. [14] considered the issues associated with the detection of multiple human classifications using a novel statistical model of weighted segmentation and a rank correlation-based feature selection approach. The proposed method was validated on six datasets based on seven performance measures. A fair comparison with the existing work was also provided which proved the significance of the proposed compared to the other techniques.

The aim of the present work is to recognize the input sequence of human actions using SIFT descriptor. The use of SIFT descriptor in HAR operation is a real challenge due to its scale invariant. It is intended to preprocess the query video sequence and reformat the SIFT features in the form that serves the HAR task, which is performed by encoding the SIFT features and then constituting the BoW database for each human action in the used dataset. HAR is then achieved by comparing the BoW of human action extracted from the query video sequence with that found in the database.

3. PROPOSED METHOD

Our proposed method depends on the concepts of multistage preparation of the material video and highlighting the most descriptive features. The use of SIFT descriptor is useful to achieve higher stable results in spite of the change in scale or pose. The following subsection explains more details.

3.1 Video Shots

A video shot is the first step for getting a sequence of image frames from the input video. Videos in a specified dataset are usually extended along 1-3 second time with low frame rate of about 10 *frame per second*. The description of human action needs a sequence of video frames (e.g., at least five frames) for completing the understanding of human action. Therefore, the given video should be sliced to a certain time interval range of 200-500 *m sec* in order to gives 5 frames (i.e., shots) for each sequence.

3.2 Erosion Filter

The erosion filter is one of the basic morphological operators (mathematical operations), which is often applied to binary images. It makes the pixels that are surrounding the dark regions darker (i.e. like reduction) [15].

The image can be symbolized by f(x) and the grayscale structuring elements by e(x), while the grayscale Erosion can be expressed as:

$$(f \ominus e)_{(x,y)} = f(x+s, y+p) - e(s,p) \quad \ldots (1)$$

where $\ominus$ denotes to the subtraction process between the image f(x, y) and erosion kernel e(s, p), s and p are indices referring to the position of each element in the kernel. In other words, the erosion filter works as, for each pixel, taking a minimum value of the surrounding neighbors in squared kernel of a size of 3×3, 5×5, or 7×7 *pixels*. The noise is almost removed in such process, and the contours of objects in the image are quickly determined, such that the contour can be achieved by subtracting the eroded image from the original one.

3.3 Otsu Thresholding

The process of converting a color image into monochrome is common in the area of image processing. Otsu's thresholding method is used to binarize the eroded video frame. The application of this technique involves repeating over all the potential values of threshold and computing an amount of pixel levels spread on both threshold sides, i.e. all pixels take place in either background or foreground. It is important to obtain the threshold number whereas the total of background and foreground spreads is at its lowest [16].

$$A = W \times H \quad \ldots (2)$$

$$w_f = \sum_{i \epsilon f} x_i \,/A \quad \ldots (3)$$

$$\mu_f = \left[\sum_{i \epsilon f} (i \times x_i)^2\right] / \left[\sum_{i \epsilon f} x_i\right] \quad \ldots (4)$$

$$\sigma_f^2 = \left[\sum_{i \epsilon f} (i \times \mu_f)^2 \times x_i\right] / \left[\sum_{i \epsilon f} x_i\right] \quad \ldots (5)$$

$$w_b = \sum_{i \epsilon b} x_i / A \quad \ldots (6)$$

$$\mu_b = [\sum_{i \epsilon b} (i \times x_i)^2] / [\sum_{i \epsilon b} x_i] \quad \ldots (7)$$

$$\sigma_b^2 = [\sum_{i \epsilon b} (i \times \mu_b)^2 \times x_i] / [\sum_{i \epsilon b} x_i] \quad \ldots (8)$$

where each of *W*, *H*, and *A* are the width, height, and area of the image, respectively, f is the foreground, b is the background, $\boldsymbol{x}$ is the occurrence of ith pixel, and w_f and w_b are weights of foreground and background pixels, respectively. While, both of μ_f and μ_b are the means of foreground and background pixels, respectively, and σ_f^2 and σ_b^2 represent the variances of the foreground and background pixels.

$$\sigma_{within}^2 = w_f \sigma_f^2 + w_b \sigma_b^2 \quad \ldots (9)$$

$$\text{Otsu Threshold} = min\left\{\sigma_{within}^2\right\} \quad \ldots (10)$$

3.4 Region of Interest (RoI)

For detecting specific points or areas in digital images, a mathematical process, which is blob detection, is used. A point or an area that holds distinguished changes with its surrounding is called Blob. The Blob is a region that can be either brighter or darker than the neighborhood in a video or an image. Therefore, objects, humans, and even marks in the image may appear as Blob.

Blob detectors can be categorized into two types of techniques. The first type is called the differential techniques, which is a derivative function based on the blob position. While, the second is the finding of the local minima or maxima of the function. Theoretically, the second technique can provide complete information about all regions. Therefore, it can be used to extract the RoI for further processing [17].

$$\forall\, p \in I = \begin{cases} 0 & p \in b \\ 255 & p \in f \end{cases} \quad \ldots (11)$$

$$\{\forall\, p \in f\ : p\ \longrightarrow l = l + 1\ \} \quad \ldots (12)$$

$$\forall\, p_{\ labeled} = \min(Neighbor\, p_{\ labeled}) \quad \ldots (13)$$

$$\forall\ p_{labeled} \in object: get\ (x_{min}, y_{min}, x_{max}, y_{max})\ \forall\ object \quad \ldots (14)$$

where *p* represents the i[th] pixel of the image *I*, *f* and *b* are foreground, and background respectively, *l* is label number, which begins from 0, $p_{\ labeled}$ is the pixel that takes the least label number among its neighbors; if the *object* has the same label number, it will be a candidate to get its coordinate (i.e. (x_{min}, y_{min}, x_{max}, y_{max})).

3.5 SIFT Descriptor

Scale Invariant Feature Transform (SIFT) is the most popular algorithm in computer vision. It used to describe and detect local features in the images.

Using this algorithm, the interested points for any object in an image can be extracted to give the description features of that object. This algorithm was patented and published by David Lowe in 2004 [18]. In this stage, each object obtained from a previous step is entered to a SIFT descriptor, where the SIFT algorithm consists of four parameters:

- Keypoints Detection
- Keypoints Refinement
- Orientation Assignment
- Keypoints Description

3.5.1 Keypoints Detection

Keypoints are detected by comparing the central pixel of the current keypoint with its surrounding 8-neighbors; if it is larger than the neighbors it becomes maxima whereas if it is smaller the neighbors become minima.

3.5.2 Keypoints Refinement

All keypoints obtained in the detection stage undergo a refinement test in order to sort the useful keypoints and eliminate those with low contrast.

A. Refining Keypoints and Discarding Low Contrast

When a candidate keypoint has been located by comparison with its neighbors, the next step is performing a detailed fit to the nearby data for location, i.e. peak magnitude and edge response. Therefore, this information permits a point to be excluded since it had low contrast and noise sensitivity, or it is weakly centralized along an edge. Thus, to get more accurate results, all the peaks have to be refined. Taylor series expansion, which is given in equation (15), is used on candidate keypoints to get more a accurate location of extrema, and if the value at this extremum is less than a threshold value (0.03), it is rejected.

$$D(x) = D + \frac{\partial D^T}{\partial \mathbf{x}}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\frac{\partial^2 D}{\partial \mathbf{x}^2}\mathbf{x} \quad \ldots (15)$$

where *D* is [evaluated] at a sample candidate point, in which $\mathbf{x} = (x, y, \sigma)^T$ is the offset from this point. The first derivative is the gradient (*G*), while the second derivative represents the 3x3 Hessian matrix (*H*), and both G and H are computed as pixel differences. For i = 0 to 4, and for each keypoint, the 3D Gradient (G) is calculated as follows:

$$G = \begin{bmatrix} \frac{(g_{i+1}(x,y) - g_i(x,y))}{2} \\ \frac{(g_i(x+1,y) - g_i(x-1,y))}{2} \\ \frac{(g_i(x,y+1) - g_i(x,y-1))}{2} \end{bmatrix} \quad \ldots (16)$$

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{13} & h_{23} & h_{33} \end{bmatrix} \quad \ldots (17)$$

where the 3×3 Hessian Matrix elements are given as follows:

$$\left.\begin{aligned} h_{11} &= g_{i+1}(x,y) + g_{i-1}(x,y) - 2 \times g_i(x,y) \\ h_{22} &= g_i(x+1,y) + g_i(x-1,y) - 2 \times g_i(x,y) \\ h_{33} &= g_i(x,y+1) + g_i(x,y-1) - 2 \times g_i(x,y) \\ h_{12} &= (g_{i+1}(x+1,y) - g_{i+1}(x-1,y) - g_{i-1}(x+1,y) + g_{i-1}(x-1,y))/4 \\ h_{13} &= (g_{i+1}(x,y+1) - g_{i+1}(x,y-1) - g_{i-1}(x,y+1) + g_{i-1}(x,y-1))/4 \\ h_{23} &= (g_i(x+1,y+1) - g_i(x+1,y) - g_i(x-1,y+1) + g_i(x-1,y-1))/4 \end{aligned}\right\} \quad \ldots (18)$$

The position of the extrema $\hat{x}$ can be defined by determining the first derivative of that function with reference to x, and equal it to zero. This gives:

$$\hat{\mathbf{x}} = -\frac{\partial^2 D^{-1}}{\partial \mathbf{x}^2}\frac{\partial D}{\partial \mathbf{x}} \quad \ldots (19)$$

When the offset value $\hat{x}$ is greater than 0.5 in any range, it is not extreme but approaching to another test point. In such a situation, the test point was modified and the incorporation was achieved instead about this point. Then the last offset value $\hat{x}$ is added to the position of its sample point to get the incorporation approximated for the position of the extreme. The D ($\hat{x}$) represents the peak extremum in the differences of Gaussians values, and it is suitable for discarding unsteady peaks that have low contrast. It can be attained by the compensation of equation (19) into equation (15), yielding the following relation:

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2}\frac{\partial^2 D^T}{\partial \mathbf{x}}\hat{\mathbf{x}} \quad \ldots (20)$$

If any absolute values of D ($\hat{x}$) are less than 0.03, they are rejected as weak extremes (i.e. low contrast points).

B. Discarding Keypoint Edges

This step is a process of discarding keypoints that lie on the edges, as the edges are highly responsive to the differences of Gaussians filter. All weak extrema have a high maximal curvature at edge but a very low minimal one in the vertical direction.

The scale and position located of these basis curvatures for each keypoint are determined by a 2×2 Hessian Matrix as follows:

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \quad \ldots (21)$$

The derivatives D_{xx}, D_{xy}, and D_{yy} are approximated by taking the variances of adjacent sample points as follows:

$$D_{xx} = g_i(x+1, y) + g_i(x-1, y) - 2 \times g_i(x, y) \quad \ldots (22)$$

$$D_{yy} = g_i(x, y+1) + g_i(x, y-1) - 2 \times g_i(x, y) \quad \ldots (23)$$

$$D_{xy} = \big(g_i(x+1, y+1) - g_i(x+1, y-1) - g_i(x-1, y+1) + g_i(x-1, y-1)\big)/4 \quad \ldots (24)$$

The Eigenvalues of H are relative to the principle curvatures of D. The Hessian matrix could prevent the calculation of the Eigenvalues, as their ratio is just involved. Let α be the greatest of the Eigenvalues and β the lowest one. Therefore, one can evaluate the total of the Eigenvalues from the Trace (T_r) of H and their result from the Determinant (Det) as follows:

$$T_r(H) = D_{xx} + D_{yy} = \alpha + \beta \quad \ldots (25)$$

$$\text{Det}(H) = D_{xx}D_{yy} - \left(D_{xy}\right)^2 = \alpha\beta \quad \ldots (26)$$

Let r be the proportion between the greatest and lowest Eigenvalues; $\alpha = r\beta$. It leads to:

$$\frac{T_r(H)^2}{\text{Det}(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\alpha+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r} \quad \ldots (27)$$

Equation (27) is only based on the ratio of the Eigenvalues rather than their individual values. The amount of $(r + 1)^2 / r$ has a lowest value at the time that the two Eigenvalues are the same, and it increases with increasing (r) value. Hence, one can verify whether the proportion of the principal curvatures is less than a specific threshold (r) through:

$$\frac{\text{Tr(H)}^2}{\text{Det(H)}} < \frac{(r+1)^2}{r} \quad \ldots (28)$$

The use of above the condition is very important to refine keypoints under consideration, with less than 20 floating point operations required to test each location. The experiments showed that the proper value of r is 10, which leads to eliminate the keypoints of principle curvatures of greater than 10.

3.5.3 Orientation Assignment

The orientation of each keypoint is assumed to be dependent on local image properties. The descriptor of a keypoint can be characterized proportional with this orientation in order to obtain images with invariant rotation. The keypoint scale is used to choose the average refined image with the nearest scale, where the calculations are achieved in a scale of invariant manner. The histogram of

orientation is constructed from the orientations of gradient at all test points inside a rounded window about that keypoint. Each of test points that are put into the histogram is measured by its gradient

magnitude and by the Gaussian weight rounded window with a triple scale of that keypoint. Gaussian window aims to prevent unexpected modifications in the descriptor with less variations in the location of that window, and to assign some gradients that are distant away from the center of descriptor, where these have the greatest impacts by misregistration faults [18].

3.5.4 Keypoints Description

After each image orientation, scale, and location is assigned to each keypoint, these parameters require a 2D coordinated system to describe the local image region. The following step is calculating a regional image area descriptor which is very characteristic but is as invariable as potential to the rest of the parameters, including the viewpoint or illumination changes. Thus, a descriptor of the regional image for each keypoint is constructed relying on the orientations and magnitudes of the image gradients in the district of the keypoint, where a 16×16 window around a keypoint is partitioned to sub-regions of 4×4 that contain the gradient magnitude and orientation.

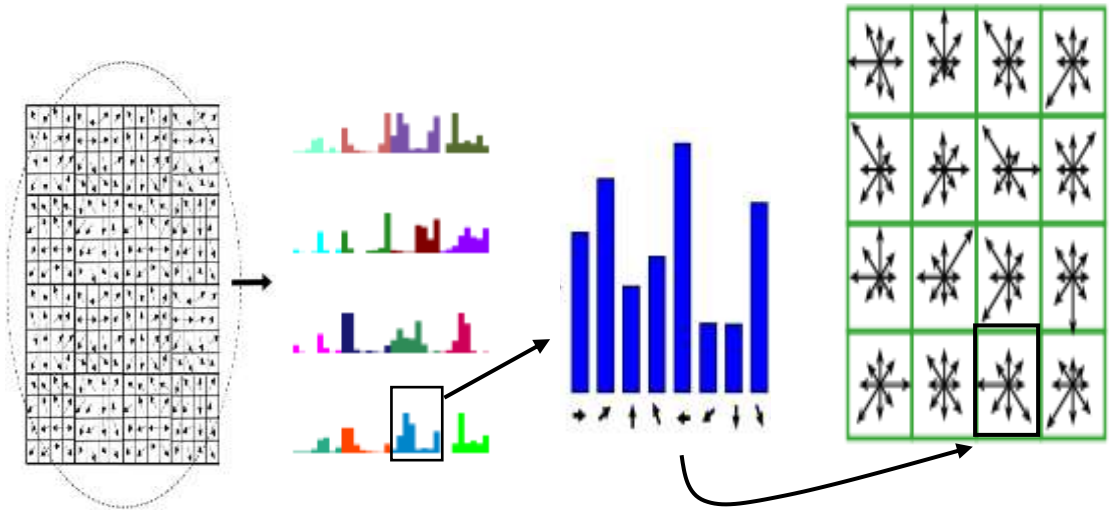


Figure 2- Keypoint descriptors estimation.

3.6 Enrollment Phase

The enrollment phase is concerned with collecting the comparable features of each sample in the dataset to produce a BoW database. The enrollment includes all features of each keypoint obtained from the descriptor array of the contributed classes. The process of creating BoW includes combining all features that belong to a specific class in one codeword and store it in a database file.

3.7 Classification Phase

The classification phase is concerned with the BoW that is extracted, based on image features from given video sequences, and then compared with those found in the database for the issuing of a corresponding action class label. The classification algorithm contains the same pre-processing and extraction features of the enrollment phase. The additional stages over the stages of the enrollment phase are the comparison-based similarity measuring and decision making. The comparison is carried out between the BoW of a given sample and its corresponding BoW of each class found in the database. The similarity measure stage uses the Euclidean distance given in:

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - y_n)^2 \quad \ldots (29)$$

where N is the length of the data sequence, x_n represents the input of data and y_n is the recreated data of n^{th} order, which is used to indicate the amount of convergence of a given video sample from those found in the database. Classification decision is made depending on the results of the similarity measurement stage. The vector of similarity scores is input to find out the most similar class, which is the class that has the greater similarity score in the vector V_R. Such that, the classification decision is made to show a message that refers to the label of the most similar class that showed the greater similarity score.

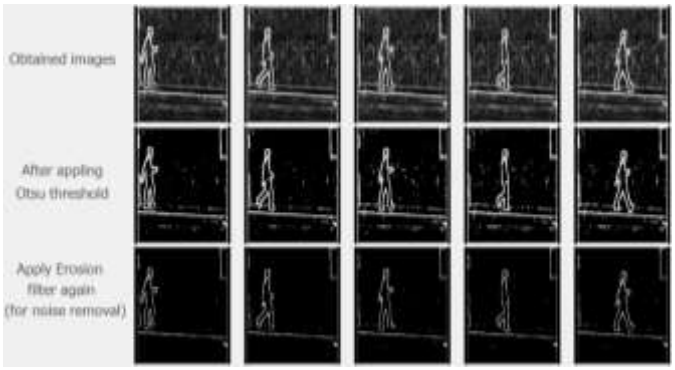


Figure 3-Applying Erosion filter.

4. RESULTS AND DISCUSSION

The dataset used is WEIZMANN [19], which is the most common dataset for video analysis of human actions. In this database, all sequences are taken with a static camera of 25 *fps* frame rate, where the sample size is reduced to the spatial resolution of 180×144 *pixels*. Each video sample in this dataset has one type of action, while the dataset holds ten action types, which include bend, jack, jump, p-jump (parallel jump), run, side, skip, walk, wave with one hand, and wave with two hands. Each action type is performed by different persons dressing diverse clothes. Figure-3 illustrates the results of the erosion filter applied on a walk video sequence. To denoise, an erosion filter was used with 3x3 kernel, as in the result shown in Figure-4, while Figure-5 shows the extracted RoI from each frame for the same video sample and class. Figure-6 presents the keypoint detection, whereas Figure -7 shows the results of keypoints refinement. Figure-8 shows the circling of each descriptor keypoint after placing RoI on the blacked frame.

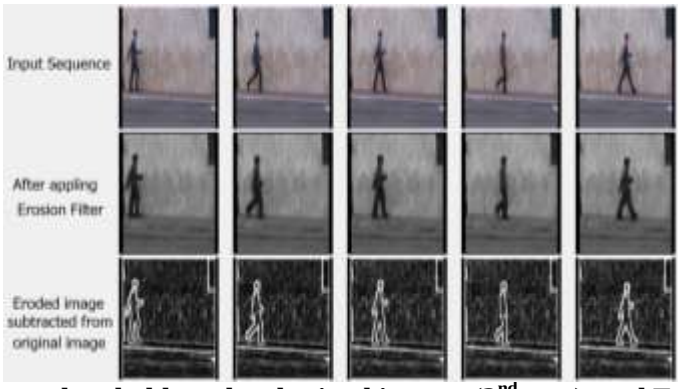
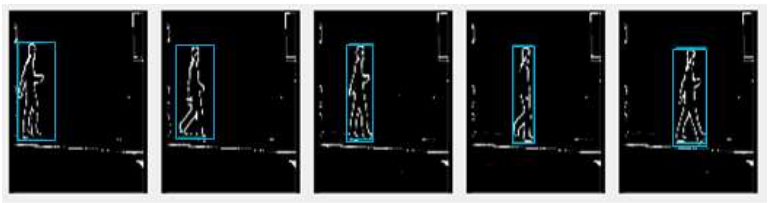


Figure 4-Applying the Otsu threshold on the obtained images (2nd row), and Erosion filter (3rd row).



(a) Surrounding box around each blob.

(b) Crop this blob as RoI.

Figure 5-Blob detection for getting RoI.

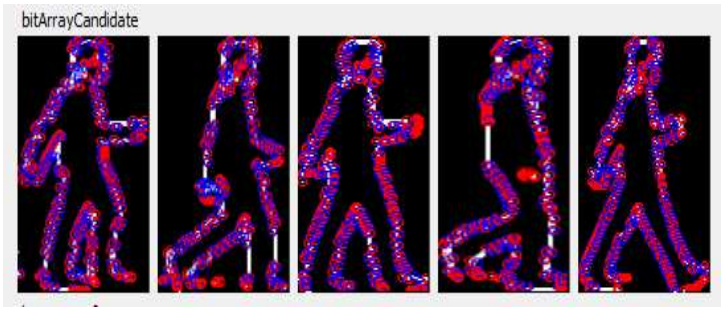


Figure 6-Candidate keypoints: red for maxima and blue for minima.

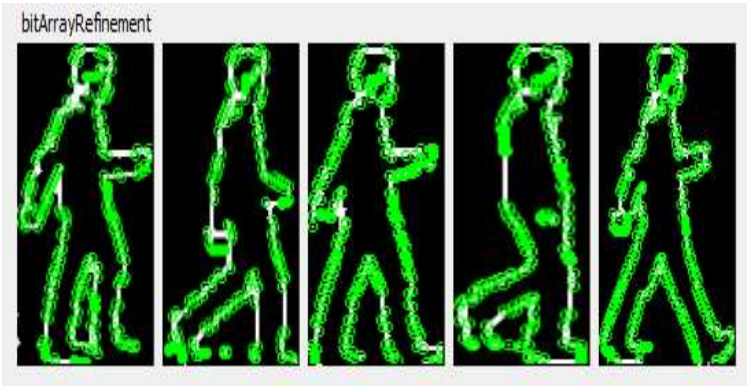


Figure 7-Result of keypoint refinement process applied on same test video sequence sample.

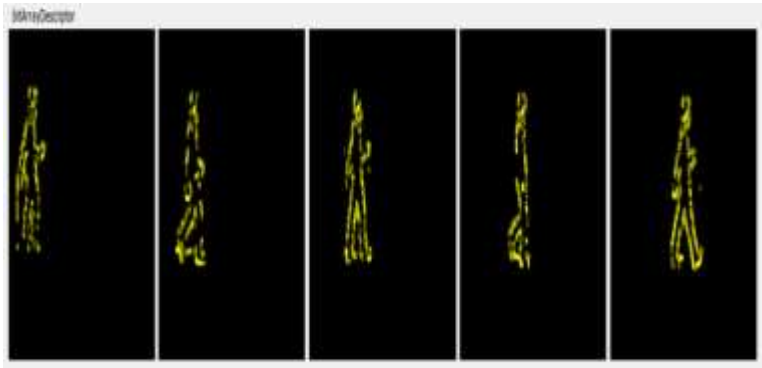


Figure 8-Descriptor keypoints after circling each keypoints.

The process of preparing BoWs is equivalent to the process of features mapping, in which each description keypoint had a feature vector that has 128 values that describe one keypoint. These values are normalized to a single value by obtaining the mean square root to their summation. Such that, the 128 values feature of specific keypoint are converted into five new values only, which are the keypoint position (x, y), keypoint deviation (σ), keypoint orientation (θ), and normalized feature vector (f). The modified description of keypoints includes only five parameters and prepared to be stored as a single code in the current stage. The combination of these codes with each other enables to create one word which is stored in a specific Bag for the purpose of comparison-based recognition process. Figure-9 shows the process of BoW constitution, while Figure-10 shows a message with maximum classes match. Table-1 presents the confusion matrix of recognition results for randomly chosen video samples from a used dataset to input the recognition phase.

X	Y	σ_c	θ key	Vector Features (128 values)
2	76	0.8	82.8091	<21 7 0 0 …… 14 16 2 0 0>

Codeword	2,76,0.8, 82.8091,455.297 Coding

Codeword	BoW
1	2,76,0.8, 82.8091,455.365
2	7,35,0.3, 27.7350,387.297
3	84,37,0.7, 74.7359,423.475
4	47,38,0.2, 46.7395,635.352
5	25,63,0.5, 49.7003,402.173
….	……..

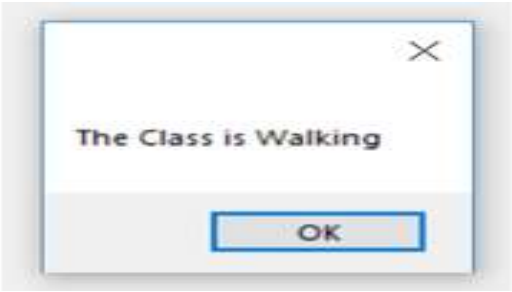
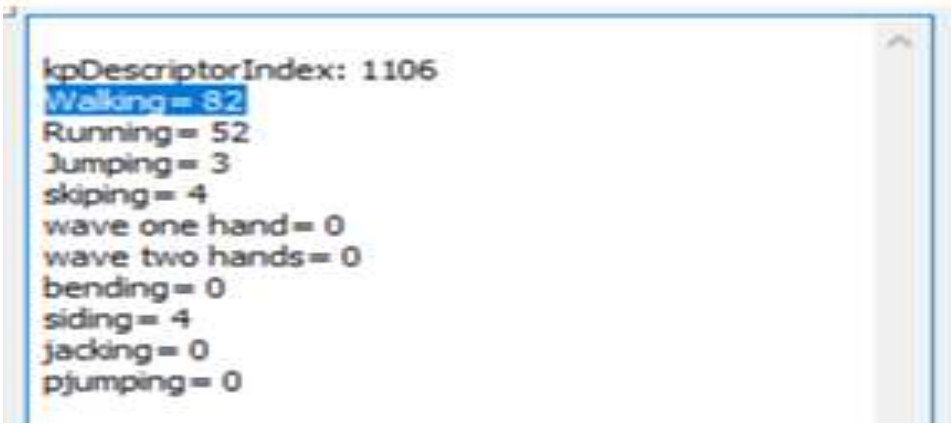


Figure 9-Feature mapping



(a) Message describes the type of action to which it belongs.
(b) The determined class is that of maximum similarity between other classes.
Figure 10: Class matching interface.

Table 1-Recognition of an action type for a randomly chosen video sample from the dataset used

Action Type	Skip	Walk	Run	Pjump	Jack	Wave 1	Bend	Wave 2	Side	Jump
Skip	0.13	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.25	0.38
Walk	0.00	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.43
Run	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.17	0.17
Pjump	0.00	0.00	0.00	0.67	0.00	0.00	0.17	0.17	0.00	0.00
Jack	0.00	0.00	0.00	0.00	0.71	0.00	0.00	0.00	0.29	0.00

Wave 1	0.00	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.00	0.17
Bend	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
Wave 2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Side	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Jump	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

3.6.5 Results Evaluation

One of the popular evaluation measures is the Receiver Operating Characteristic (ROC) curves which are a valuable visual tool for comparing models of classification. ROC curves come from the signal detection theory that was developed during World War II for the analysis of radar images. A ROC curve for a given model shows the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). For a given test set and a model, TPR is the proportion of positive (or "yes") tuples that are correctly labeled by the model; FPR is the proportion of negative (or "no") tuples that are mislabeled as positive.

True Positives (*TPs*): the positive rows which are accurately identified by a classifier.
True Negatives (*TNs*): the negative rows that are accurately identified by a classifier.
False Positives (*FPs*): the negative rows that are misidentified as "positive".
False Negatives (*FNs*): the positive rows that are misidentified as "negative".

The ROC curve is a commonly used measure in the area of discrimination. The measurements of TPR and FPR are calculated and illustrated in Table-2, which determines the system's ability to distinguish human actions for each class. Figure-11 shows the behavior of the ROC curve corresponding to the measurements given in Table-2, whereas the accuracy (A_r) and the error rate or misclassification rate (E_r) for each class are given in Table-3.

Table 2-Calculated TPR and FPR for each class.

Action Type	**TP**	**FN**	**FP**	**TN**	**P = TP + FN**	**N = FP + TN**	**Sensitivity (TPR) = TP / P**	**Specificity (TNR) = TN / N**	**FPR = FP / N**
Skip	0.13	0.88	0.00	9.00	1.00	9.00	**0.13**	1.00	**0.00**
Walk	0.57	0.43	0.00	9.00	1.00	9.00	**0.57**	1.00	**0.00**
Pjump	0.66	0.34	0.00	9.00	1.00	9.00	**0.66**	1.00	**0.00**
Run	0.66	0.34	0.25	8.75	1.00	9.00	**0.66**	0.97	**0.03**
Jack	0.71	0.29	0.00	9.00	1.00	9.00	**0.71**	1.00	**0.00**
Wave1	0.83	0.17	0.00	9.00	1.00	9.00	**0.83**	1.00	**0.00**
Wave2	1.00	0.00	0.17	8.83	1.00	9.00	**1.00**	0.98	**0.02**
Bend	1.00	0.00	0.17	8.83	1.00	9.00	**1.00**	0.98	**0.02**
Side	1.00	0.00	0.71	8.29	1.00	9.00	**1.00**	0.92	**0.08**
Jump	1.00	0.00	1.15	7.85	1.00	9.00	**1.00**	0.87	**0.13**

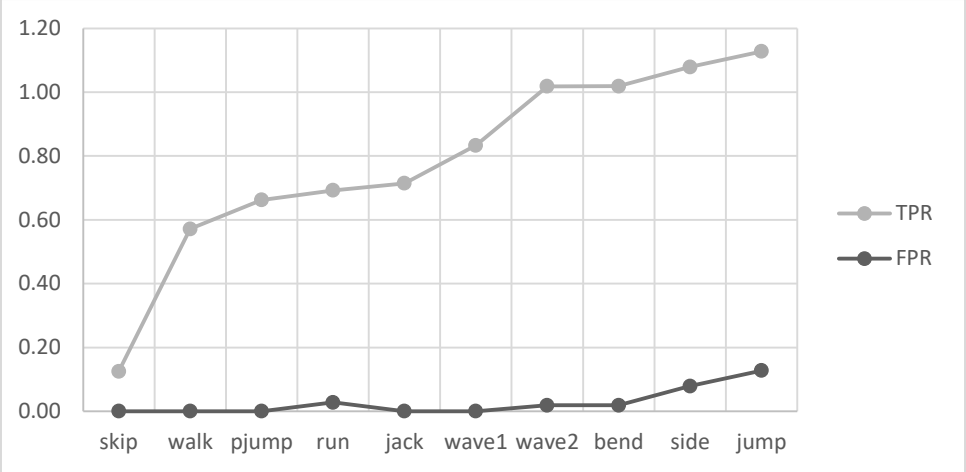


Figure 11-Show the ROC curve for proposed method applied on used dataset.

Table 3-The estimated value of A_r and E_r for each class

Action Type	A_r %	E_r%
Bend	98.3	1.7
Jack	97.14	2.86
Pjump	96.67	3.33
Wave1	98.37	1.63
Wave2	98.3	1.67
Run	94.17	5.83
Side	92.9	7.1
Skip	91.28	8.72
Jump	88.5	11.5
Walk	95.75	4.25

4. CONCLUSIONS AND FUTURE WORK

The use of Erosion filter gives a better result than the Robinson and Sobel filters in terms of edge detection as well as noise removal. Otsu threshold works as a good tool for binarizing a given image, such that the eroded image is converted into two regions, i.e. foreground and background. The use of blob detection gives a better representation to the region of interest, which showed well extracting of the human body. SIFT method shows accurate results as a feature extraction algorithm.

The refinement process leads to eliminate undesired keypoints that may confuse the description process, which is the process that leads to enhance the recognition score. The process of dimension reduction from 128 value to only five descriptors shows a better description for keypoint description. The use of BoW in the comparison-based recognition shows a higher recognition score due to less descriptor and more informatics code used in the comparison-based recognition. The use of higher power similarity measure produces more accurate recognition results. The computed TPR and FPR measures gave a true indication about the proposed HAR system, so it is exhibiting the situation of HAR system. The computed accuracy and the misclassification rate reveal the effectiveness of the system with the used dataset.

The proposed future work includes:

(a) Improving the method to be more effective, extending the algorithm to describe human actions in dynamic scenes, and then the proposed method will be used to deal with action recognition with multiple persons and tested on more challenging datasets.

(b) Applying the method on additional datasets which hold more compound actions, and examining the robustness of the method under several scales and directions.

References

1. Kong, Y. and Yun F. **2018**. Human Action Recognition and Prediction: A Survey", arXiv, *Journal of Latex Class Files*, **13**(9): September 2018.
2. Zhang, H.B., Zhang, Y.X., Zhong, B., Lei, Q., Yang, L., Du, J.X. and Chen, D.S., **2019**. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors*, **19**(5): 1005: 1-1005:20.
3. Dhamsania, C. J. and Ratanpara, T. V. **2016**. A Survey on Human Action Recognition from Videos. 2016 Online International Conference on Green Engineering and Technologies (IC-GET). pp. 1-5 IEEE. DOI:10.1109/get.2016.7916717.
4. Tsai, Chih-Fong. **2012.** "Bag-of-Words Representation in Image Annotation: A Review.". *ISRN Artificial Intelligence*. 2012 Nov 29.
5. Swarnambigai, V.K. **2014**. Action Recognition Using AMI and Support Vector Machine. *International Journal of Computer Science and Technology*, IJCST, **5**(4): Oct - Dec 2014.
6. Kazui, M., Miyoshi, M., Muramatsu, S. and Fujiyoshi, H. **2008**. Incoherent Motion Detection Using A Time-Series Gram Matrix Feature. 2008 19th International Conference on Pattern Recognition, 2008. IEEE, 1-5.
7. Wonjun K., Jaeho L., Minjin K., Daeyoung, O. and Changick, K. **2010**. "Human Action Recognition Using Ordinal Measure of Accumulated Motion", *In EURASIP Journal on Advances in Signal Processing*, 2010.
8. Song, Y., Tang, S., Zheng, Y.-T., Chua, T.-S., Zhang, Y. and Lin, S. **2010**. A Distribution Based Video Representation For Human Action Recognition. 2010 IEEE International Conference on Multimedia and Expo, 2010. IEEE, pp. 772-777.
9. [9] Eweiwi, A., Cheema, S., Thurau, C. and Bauckhage, C. **2011**. Temporal Key Poses For Human Action Recognition. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011. IEEE, pp. 1310-1317.
10. Escorcia, V. and Niebles, J. **2013**. Spatio-Temporal Human-Object Interactions For Action Recognition in Videos. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2013. Pp. 508-514.
11. Lo, S.-L. and Tsoi, A.-C. **2014**. Motion Boundary Trajectory For Human Action Recognition. Asian Conference on Computer Vision, 2014. Springer, pp. 85-98.
12. Wang, Y., Lu, Q., Wang, D., Liu, W. J. J. O. E. and Engineering, C. **2015**. Compressive Background Modeling For Foreground Extraction. Vol, 13.
13. Gaba, N., Barak, N. and Aggarwal, S. **2016**. Motion Detection, Tracking and Classification For Automated Video Surveillance. 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), 2016. IEEE, pp. 1-5.
14. Sharif, M., Khan, M. A., Zahid, F., Shah, J. H., Akram, T. J. P. A. and Applications **2019**. Human Action Recognition: A Framework of Statistical Weighted Segmentation and Rank Correlation-Based Selection. pp. 1-14.
15. Sarkar, A. R., Sanyal, G. and Majumder, S. J. I. J. O. C. A. **2013**. Hand Gesture Recognition Systems: A Survey. Vol, 71.
16. Vala, H. J., Baxi, A. J. I. J. O. A. R. I. C. E. and Technology **2013**. A review on Otsu image Segmentation Algorithm. Vol, 2, pp. 387-389.
17. Han, K. T. M. and Uyyanonvara, B. **2016**. A Survey of Blob Detection Algorithms for Biomedical Images. 2016 7th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), 2016. IEEE, pp. 57-60.
18. Lowe, D.G. **2004**. "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, **60**(2): 91–110,.
19. http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html