This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

Robust Feature-based Automated Multi-view Human Action Recognition System

Kuang-Pen Chou[1], Mukesh Prasad[2], Di Wu[2], Nabin Sharma[2], Dong-Lin Li[3], Yu-Feng Lin[1], Michael Blumenstein[2], Wen-Chieh Lin[1] and Chin-Teng Lin[2]

[1]Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan
[2]Centre for Artificial Intelligence, School of Software, FEIT, University of Technology Sydney, Australia
[3]Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan

***Abstract*—Automated human action recognition has the potential to play an important role in public security, for example, in relation to the multi-view surveillance videos taken in public places, such as train stations or airports. This paper compares three practical, reliable and generic systems for multi-view video-based human action recognition, namely the nearest neighbor classifier (NNC), Gaussian mixture model classifier (GMMC) and the nearest mean classifier (NMC). To describe the different actions performed in different views, view-invariant features are proposed to address multi-view action recognition. These features are obtained by extracting the holistic features from different temporal scales which are modeled as points of interest which represent the global spatial- temporal distribution. Experiments and cross-data testing are conducted on the KTH and WEIZMANN and MuHAVi datasets. The system does not need to be retrained when scenarios are changed which means the trained database can be applied in a wide variety of environments, such as view angle or background changes. The experiment results show that the proposed approach outperforms the existing methods on the KTH and WEIZMANN datasets.**

***Index Terms*—Multi-view video, Action Recognition, Feature Extraction, Background Subtraction, Classification, Machine Learning.**

I. INTRODUCTION

Recently, human action recognition research has brought many challenges in the areas of sports, security and personal health care systems. Automatic video analysis systems which can recognize events related to human actions are becoming necessary in different industry areas. Therefore, human action recognition has become a hot research area in computer vision and there have been many papers published on this and many real-world applications have been developed, such as searching for the structure of large video archives, gesture recognition, video indexing, and video surveillance [1-7]. Human-computer interaction, in particular, is a crucial application in action recognition research. Visual cues are a significant part of human-computer interaction to enable better communication between humans and computers, hence researchers utilize visual cues to recognize gestures and actions. Most of the recent action recognition work samples an action sequence manually before it can be recognized in a film. However, it is not practical to manually set the beginning and ending of an action sequence of the film previously. Therefore, a practical recognition system needs to be able to automatically separate many actions in an image sequence.

The current published methods for action recognition often sample an action sequence manually before it is recognized in a film [8-10]. However, it is not practical that setting the beginning and end of an action sequence of the film previously. Therefore, a practical recognition system needs to separate many actions at an image sequences automatically. Moreover, actions can be performed as different subjects such as size, posture, motion and clothing, which is still a challenging problem for several reasons, such as illumination, occlusion, shadow, camera movement or other environment changes. In addition, the actions depend on or involve objects which could add another layer of variability. As a consequence, action recognition methods often assume that the action is captured under restricted and simplified environments such as static backgrounds, non-complicated action classes and static cameras. [11-15]. In particular, frequently moving the camera to an unknown position is the main cause of view variations. Similar to observing static objects from multi-view points, the actions may appear to be different from different angles. On the other hand, a moving camera could also affect the action appearance by incorporating dynamic view changes. Therefore, an action recognition system should be robust against environment and view-point changes when capturing an action sequence.

The current approaches does not require any specific parameter tuning for data processing and it explicitly exploit spatio-temporal information at multiple temporal scales. Therefore, the proposed approach is able to capture local and global temporal information as well, for interesting points of distribution. The proposed approach labels the beginning and end of the action sequence

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

automatically. In addition, the proposed method takes advantage only of the global spatio-temporal information about where and when the points of interest are detected. Therefore, it is able to capture sequence motions and occlusions at a low computational cost. In particular, the proposed approaches use view-invariant features to address multi-view action recognition from a range of perspectives.

The key contribution of this paper can be summarized as follows:

- The proposed approach labels the beginning and end of an action sequence in a video stream automatically.
- The proposed approach is able to capture sequence motions and occlusions at a low computational cost due the detection of the points of interest.
- The proposed approach applies view-invariant features to address multi-view action recognition from different perspectives. Thus, the proposed approach is robust against view changes.

The proposed novel action recognition system is more robust against view, scale and subject variance. Fig 1 shows an overview of the proposed approach for the action recognition system. It can be separated into two parts: offline training and online testing. In offline training, feature extraction is the first stage in extracting interesting information. Secondly, the feature vectors of each image sequence are described. Thirdly, the feature vectors are quantized to reduce their dimension. Finally, these vectors are stored in the database. In online testing, the first two stages are similar to offline training. Then, using the histogram range of the database, the dimension of the feature vector is reduced. Thus, the results show which action is present in the test data. The proposed approach is evaluated using the KTH dataset [16], the WEIZMAN dataset [17] and the MuHAVi dataset [18].

The rest of the paper is organized as follows. Section II details the related work. Section III describes the datasets. Section IV and V presents the feature extraction and description. All the action recognition classifiers applied to different datasets are discussed in Section VI. Section VII introduces the experiments and the results. Section VIII suggests potential research opportunities and provides a conclusion.

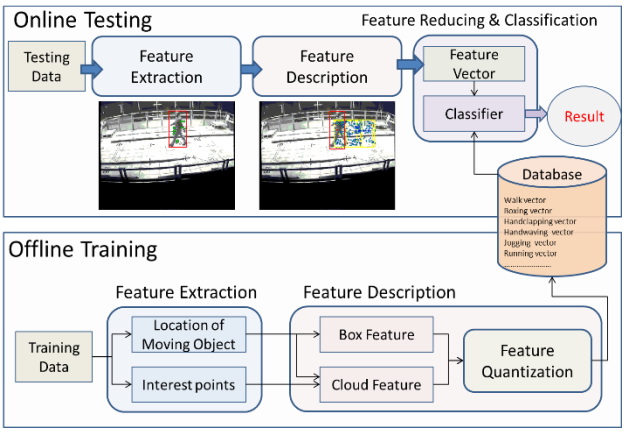


Fig. 1: System Overview

II. RELATED WORK

In the early stages of action recognition research, the techniques were based on optical flow [19, 20], tracking [21-24] and a spatio-temporal shape template [17, 25, 26]. The computation of optical flow helps to construct action templates for flow and tracking-based approaches. However, at the boundary of the segmented human body, the features are more sensitive to noise, which are extracted from the flow templates. The action recognition problem is treated as 3D object recognition by spatio-temporal shape template approaches. These approaches require the extraction of highly detailed silhouettes, which may not be possible when there is real-world noisy video input. Further, a recognition rate with 100% accuracy has been demonstrated on the WEIZMAN dataset [17], however, these approaches do not work properly on a dataset which contains noise such as the KTH dataset [16]. The KTH dataset contains noises such as low resolution, zooming, and camera movement, which makes it impossible to extract a clean silhouette. The spatio-temporal interest point-based approaches have become increasingly popular to address this problem. Further, the 2D SIFT descriptors [27] are extended to 3D with the addition of dimension to the histogram orientation by Scovanner et al. [28]. Due to the encoded temporal information, the extended 3D descriptors perform better than the 2D descriptors in action recognition. Furthermore, Willems et al. [29] proposed the spatio-temporal domain which is an extension of the SURF descriptor. Schuldt et al. [16] and Dollar et al. [30] described sparse spatio-temporal features to deal with the complexity of human action recognition [18, 31]. Schuldt et al. [16] proposed the representation of action using 3D spatio-temporal interest points captured from video frames. Schuldt also produced a histogram of informative words for

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

each action adopting the codebook and bag-of-words (BOW) approach.

A dictionary of prototypes or video-words can be formed based on the clustering of the detected points of interest. Similarly, Dollar et al. [30] introduced a multi-dimensional linear filter detector which is able to detect denser points of interest. The BOW approach was applied but it took sparser sampling of the points of interest. Niebles and Fei-Fei [32] introduced a hierarchical model which can be characterized as a constellation of bags-of-features to improve the performance. The approaches [30, 32] represent BOW features, which are adopted successfully for 2D object categorization and recognition. The BOW features are robust against noises, camera movements and low resolution datasets compared with object tracking and shape-based approaches. Moreover, these approaches mainly focus on individual local space time descriptors rather than global space time descriptors.

However, the early work did not consider noise. In recent years, researchers have applied different new methods to tackle the challenges from noise in the human action recognition area, such as camera in-variation, camera motion and occlusion. Most of the early work assumes that the action is captured from a static viewpoint without any camera movement. However, the patterns of human actions appear to be different from different angles. A person's gestures and their location vary according to each camera angle. Some of the approaches train a single classifier for all viewpoints or a set of classifiers where each classifier deals with one viewpoint [33, 34]. However, these approaches only extend the system from a single viewpoint to a multi-view dataset. Therefore, the performance only depends on the extracted features and the trained classifiers. Lu et al. [35] introduced motion history and motion energy images to observe the additional action features in the images. This approach may disrupt the background of the image especially if there is more than one person in the image. In order to obtain accurate multi-view action representations, researchers proposed some models to generate 3D or 2D body gestures through the multi-view datasets. The human body can be distinguished into several parts, and action recognition depends on the features extracted from the different body parts. Kumar and Madhavi [11] used an envelope shape to represent the human body and model the action recognition classifier.

The aforementioned approaches have difficulty ensuring the performance of the classifier when the viewpoint or environment changes. However, this paper introduces robust features to address multi-view action recognition from different perspectives and view changes as well.

III. DATASETS

The KTH Royal Institute of Technology created a dataset named the KTH Dataset [16] in 2004. It was the largest human sequence action dataset in video with different scenarios and the most popular dataset at that time, achieving a milestone in the computer vision research area. The KTH dataset includes six action classes, these being boxing, hand clapping, hand waving, walking, jogging and running. Each class is performed by twenty-five people in four different scenarios (outdoor actions, outdoor actions with zoom, outdoor actions with different clothing and indoor actions). There are a total 25x6x 4 = 600 video files in the dataset and each video only contains one person performing a single action as shown in Fig 2. The resolution and length of each video is 160 x 120 and ten to fifteen seconds respectively captured at twenty-five FPS.

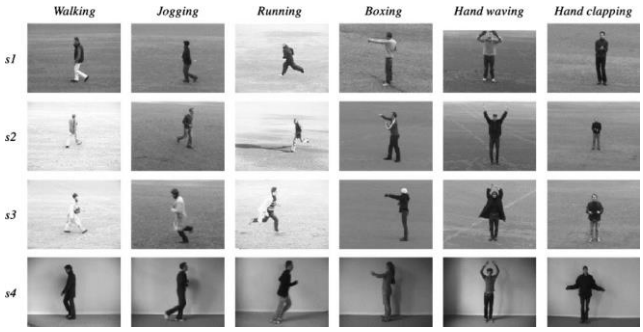


Fig. 2: Examples of KTH dataset. The four different scenarios are outdoor actions (s1), outdoor actions with zooming (s2), outdoor actions with different clothes (s3) and indoor actions (s4).

The Weizmann Institute of Science created a dataset named the WEIZMANN Dataset [17] in 2005 comprising 90 low resolution (180 x 144) videos involving nine different subjects, each of whom performs 10 basic actions, as shown in Fig 3.

Kingston University collected a large multi-view human action dataset named the MuHAvi (Multicamera Human Action Video) dataset in 2010 [18]. It comprises multi-view videos of 17 different actions performed several times by 14 people in a designated action area and is captured from different angles and distances by eight cameras. The resolution of the dataset is 720 x 576 pixels and it is captured in complex backgrounds and varying lighting conditions. The eight cameras are positioned on different sides and corners on a rectangular platform, as shown in Fig 4 and Fig 5 shows six example frames from this dataset.

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

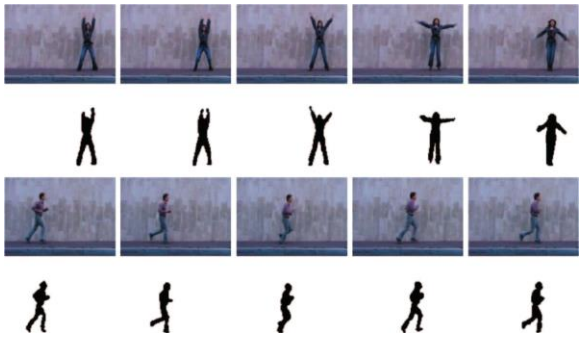


Fig. 3: Examples of Weizmann dataset includes extracted silhouettes

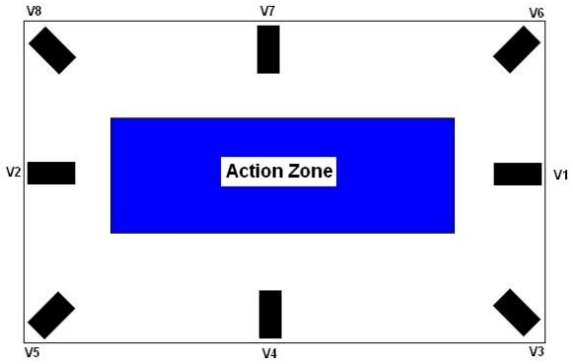


Fig. 4: Location of cameras at four sides and four corners from MuHAVi dataset

Fig. 5: Examples of MuHAVi dataset

IV. FEATURE EXTRACTION

This section describes feature extraction which includes information on moving object extraction and points of interest extraction. The details of the extraction of moving objects and interest points are shown in Fig 6.

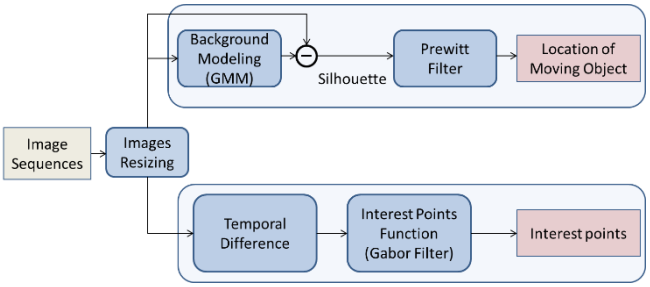


Fig. 6: Overview of feature extraction

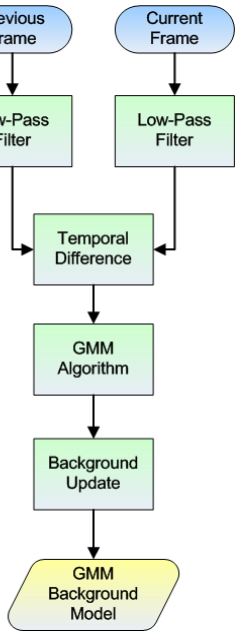


Fig. 7: GMM background model construction

A. *Moving Object Localization*

In action recognition, detecting and segmenting the foreground object without the noise produced by camera movements, zoom, shadows etc. is difficult. To do this, the model can be divided into the following steps. Firstly, the Gaussian mixture model (GMM) is used [36] to construct the background and obtain the silhouette by background subtraction. Secondly, the Prewitt edge detector [37] can be used to segment the objects from the foreground. The GMM is a common and robust method in background construction. For the purpose of action recognition in a complex scene condition, the GMM is used to build the background image. It is described as follows.

The intensity of each pixel varies in a small interval except in the region of foreground objects. It is appropriate to use a Gaussian model to construct the background image. However, in many surveillance videos, if there are waving leaves, sparking light, etc.

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

Some background pixels vary in several specific intervals. In other words, using two, three or more Gaussian distributions to model a pixel will obtain better performance. The flow chart of the GMM background construction is presented in Fig 7.

Firstly, a low-pass filter is used to reduce the noise. The GMM method models the intensity of each pixel with K Gaussian distributions. The probability that a certain pixel has a value of X_t at time t can be written as:

$$P(X_t) = \sum_{k=1}^{K} \omega_{k,t} \cdot \eta(X_t, \mu_{k,t}, \sum_{k,t}) \quad (1)$$

where K is the number of distributions that are used, $\omega_{k,t}$ represents the weight of k-th Gaussian in the mixture at time t, $\mu_{k,t}$ is the mean of k-th Gaussian in the mixture at time t, $\Sigma_{k,t}$ is the covariance matrix of the k-th Gaussian in the mixture at time t, and η is a Gaussian probability density function shown in Eq. 2.

$$\eta(X_t, \mu_t, \Sigma_t) = \frac{1}{(2\pi)^{n/2} |\Sigma_t|^{1/2}} \exp\{-\frac{1}{2}(X_t - \mu_t)^T \Sigma_t^{-1} (X_t - \mu_t)\} \quad (2)$$

where n is the dimension of data. In order to simplify the computation, it is assumed that each channel of data is independent and has the same variance, and it can then be assumed that the covariance matrix is as shown Eq. 3:

$$\sum_{k,t} = \sigma_k^2 \mathrm{I} \quad (3)$$

Temporal difference is applied to extract the possible back- ground regions, and update the pixels inside these regions. Then, we sort Gaussian distributions by the value of ω/σ, and choose the first B distributions to be the background model, i.e. shown as Eq. 4:

$$B = \arg\min_b (\sum_{k=1}^{b} \omega_{k,t} > T) \quad (4)$$

When a new pixel is imported (intensity is X_{t+1}), it will be checked against the K distributions in turn. If the probability value is within Eq. 5 standard deviations, this pixel is considered as background. Then, weight, mean, variance is updated using Eq. 5, 6, 7:

$$\omega_{k,t+1} = (1-\alpha)\omega_{k,t} + \alpha(M_{k,t+1}) \quad (5)$$

$$\mu_{t+1} = (1-\rho)\mu_t + \rho X_{t+1} \quad (6)$$

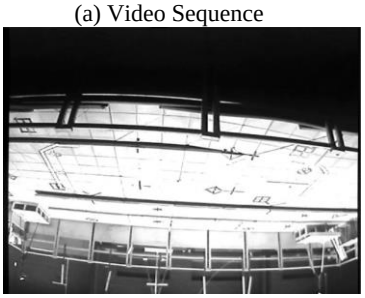
$$\sigma_{t+1}^2 = (1-\rho)\sigma_t^2 + \rho(X_{t+1} - \mu_{t+1})^T (X_{t+1} - \mu_{t+1}) \quad (7)$$

where α is the learning rate, $M_{k,t+1}$ is 1 for the model which matched and 0 for the remaining models. Eq. 8 shows the second learning rate ρ.

$$\rho = \alpha\eta(X_{t+1} | \mu_{k,t}, \sigma_{k,t}) \quad (8)$$

In addition, the remaining Gaussians only update the weight. If no distributions are matched, then the mean, variance and weight of the last distribution are replaced by X_{t+1}, a high variance and a low weight value, respectively. Fig 8 shows the background image constructed by GMM. Fig 9 shows the silhouette obtained by background subtraction. In Fig 10(a), using the edge detector to detect the location of a moving object from foreground image. In addition, a bounding box is used to indicate the presence of a foreground subject at a particular area in Fig 10(b).

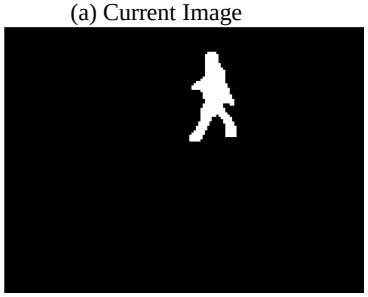
(a) Video Sequence



(b) GMM Background Image

Fig. 8: Background image construction by GMM

(a) Current Image

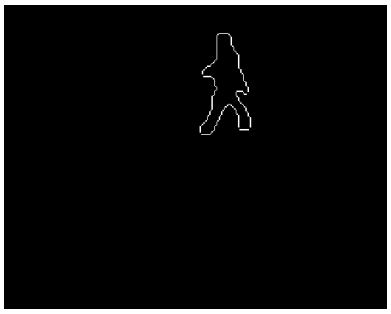


(b) Silhouette

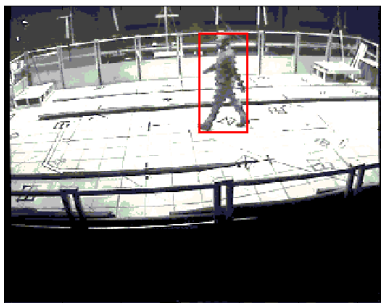
Fig. 9: Silhouette obtained by background subtraction

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access



(a) Location of Moving object



(b) Bounding Box

Fig. 10: Moving object obtained by Prewitt filter

B. Extraction of Points of Interest

The actions performed by the person should be shown in the bounding boxes. For instance, the bounding box must be located around the hands when the person performs the action "boxing". Thus, Bregonzio et al. [38] proposed a detector to capture spatio-temporal information from the bounding boxes. More specifically, the detector works in two steps: firstly, the frame differences are monitored based on the focus of the attention and detection of the region of interest. Secondly, 2D Gabor filters of different orientations are used to filter the regions of interest. These two steps give a combined filter response based on both the spatial and temporal domains.

Points of interest are local spatio-temporal features which can be considered as salient or descriptive of the action in the frames. In Dollar's method [30], the Gabor filter is used to detect intensity variations in the temporal domain. In addition, the detected points of interest correspond to local 3D peaches that represent complex actions. To be more specific, the response of the Gabor filter is given as:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \quad (9)$$

Where the Gaussian smoothing kernel can be represented as $g(x,y:\sigma)$ and can be applied in the spatial domain. h_{ev} and h_{od} are the 1D Gabor filters worked on the temporal domain which can be defined as:

$$h_{ev}(t;\ \tau,\omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2} \quad (10)$$

$$h_{od}(t;\ \tau,\omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2} \quad (11)$$

By setting the $\omega = 4/\tau$, τ and σ are the two free parameters which control the space and time scales of the detector. However, the Dollar detector has four drawbacks:

(1) The pure translational motions are ignored by the method;
(2) False detection occurs easily because of the noise in the video, which is because the approach uses local information within a small region;
(3) The approach tends to generate a spurious detection background area surrounding object boundary;
(4) The detection approach is weakened when there is slow object motion, slight camera movement or zoom.

To overcome these four problems of the Dollar detector, the detector proposed by Bregonzio et al. [38] can be utilized which proposes different filters for detecting undergoing complex motions from salient space-time local areas and capture spatio-temporal information from the bounding boxes.

The Gabor filter is a linear filter which is widely used for edge detection in image processing, and the frequency and orientation representations are similar to the human visual system. In addition, it is particularly suitable for the representation and discrimination of texture. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave. Therefore, the 2D Gabor filter has two parts, the first part $s(x, y)$ is the carrier, which represents the real part of a complex sinusoid:

$$s(x,y) = cos[2\pi(\mu_0 x + \upsilon_0 y) + \theta_i] \quad (12)$$

where μ_0 and υ_0 are the spatial frequencies of the sinusoid controlling the scale of the filter and θ_i defines the orientation of the filter. In the experiments, the 2D Gabor filters contain 5 different orientations, $\theta_i = 1, ..., 5 = \{0^\circ, 22.5^\circ, 45^\circ, 67.5^\circ, 90^\circ\}$ which shown in Fig 11.

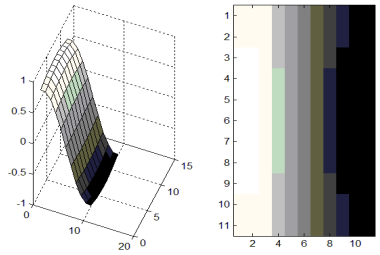
The second part of the filter $G(x, y)$ called the envelope represents a 2D Gaussian-shaped function:

$$G(x,y) = exp(\frac{x^2+y^2}{2\rho^2}) \quad (13)$$

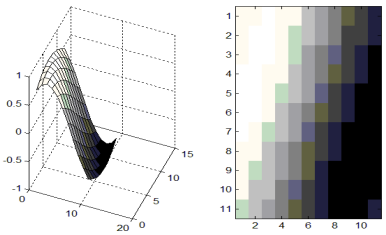
where the width of $G(x, y)$ is controlled by the parameter ρ and $\mu_0 = \upsilon_0 = \frac{1}{2\rho}$. Therefore, ρ is the only parameter control- ling the scale, which is set to 11 pixels in the experiments. By setting the threshold, the points of interest can be obtained after convolving the bounding boxes with 2D Gabor filters. Local and distinctive properties of human actions can be represented by using points of interest. Fig 12 shows the results of the point of interest detection using the MuHAVi dataset.

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

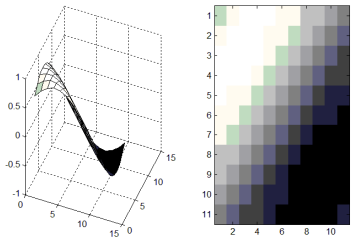
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access



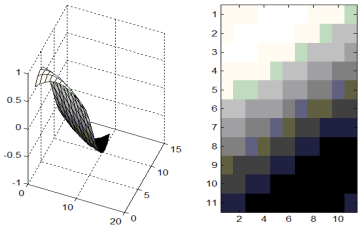
(a)



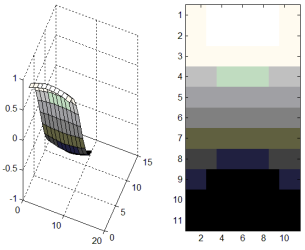
(b)



(c)



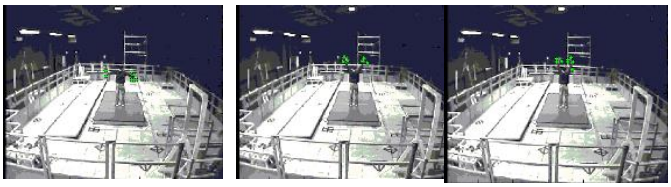
(d)



(e)

Fig. 11: Examples of the 2D Gabor filters oriented along (a) 0º, (b) 22:5º, (c) 45º, d) 67:5º and (e) 90º

(a) Walk



(b) Wave Arms

(c) Punch

Fig. 12: Results of interest point detection

V. FEATURE DESCRIPTION

This section introduces the feature vectors described by the location of moving objects and points of interest discussed in the previous chapter. Section IV.A and section IV.B illustrate the box features and cloud features. Moreover, section IV.C, describes the quantization for reducing the dimension of the feature vectors.

A. Box Feature

The first set of features is global and holistic and is concerned with the shape and speed of the foreground object. Once the object is segmented from the detected foreground area by the Prewitt edge detector [37], two features are considered: B_t^r measuring the ratio of the object height and width, B_t^{Sp} and measuring the absolute speed of the object which is normalized by the height of the object for scale invariance. Each image frame I_t has one B_t^r and one B_t^{Sp} feature.

B. Cloud Feature

Spatial information, such as human pose information, can be preserved by the detected points of interest. Moreover, the frames have a temporal dependency between each other, and in order to use

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

such information, the points of interest extracted from a set of consecutive frames are able to accumulate and form a point cloud [38]. Thus, the points of interest could represent both the spatial and temporal information for human actions.

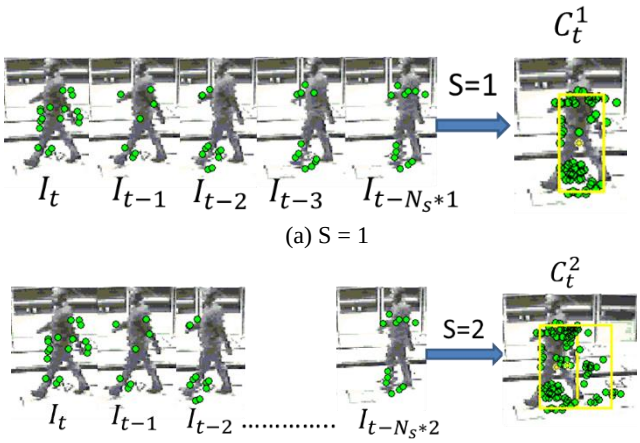
For an action video sequence, *A* contains *T* frames, which can be represented as $A = [I_1, ..., I_t, ..., I_T]$, where I_t is the *t*-th frame of the video. Then, the I_t *is set* as the current frame and the N_s as the size of a temporal scale. The sets of the past *K* cumulative scales can be defined as $[I_{t-N_S}, ..., I_t]$, $[I_{t-2\times N_S}, ...I_t]$, ...,$[I_{t-K\times N_S}, ..., I_t]$. Thus, for the specific frame I_t, there are a set of *K* interest point clouds where different temporal scales are formed. As shown in Fig 13, the clouds can be represented as $[C^1, ..., C^S, ..., C^K]$. To be more specific, by accumulating the detected points of interest over the past *S* x N_S frames, the cloud of the *s*-th scale can be built. Fig 14 shows examples of the interesting point clouds extracted from the MuHAVi dataset. It shows that different actions represent point clouds of interest which are of a different shape, relative location and distribution.

Therefore, the second set of features are called cloud features. The cloud features are scale dependent and are extracted from the point clouds of interest with different scales. Eight features are computed from the *s*-th scale cloud. The representation of the *s*-th scale cloud is as follows:

$$[C_s^r, C_s^{Sp}, C_s^{Vd}, C_s^{Hd}, C_s^{Hr}, C_s^{Wr}] \quad (14)$$

where C_s^r is the height and width ratio of the cloud. C_s^{Sp} is the absolute speed of the cloud. C_s^{Vd} and C_s^{Hd} measure the spatial relationship between the cloud and the detected object area. Specifically, $C_s^{Vd}$is the vertical distance between the geometrical centroid of the object area and the cloud, and C_s^{Hd} is the horizontal distance between the geometrical centroid of the object area and the cloud. C_s^{Hr} and $C_s^{Wr}$are the height ratio and width ratio between the object area and the cloud respectively. Overall, the six features can be put into two categories: C_s^r and C_s^{Sp} measure the shape and speed of cloud itself; the rest four features capture the relative shape and location information between the object and the cloud areas

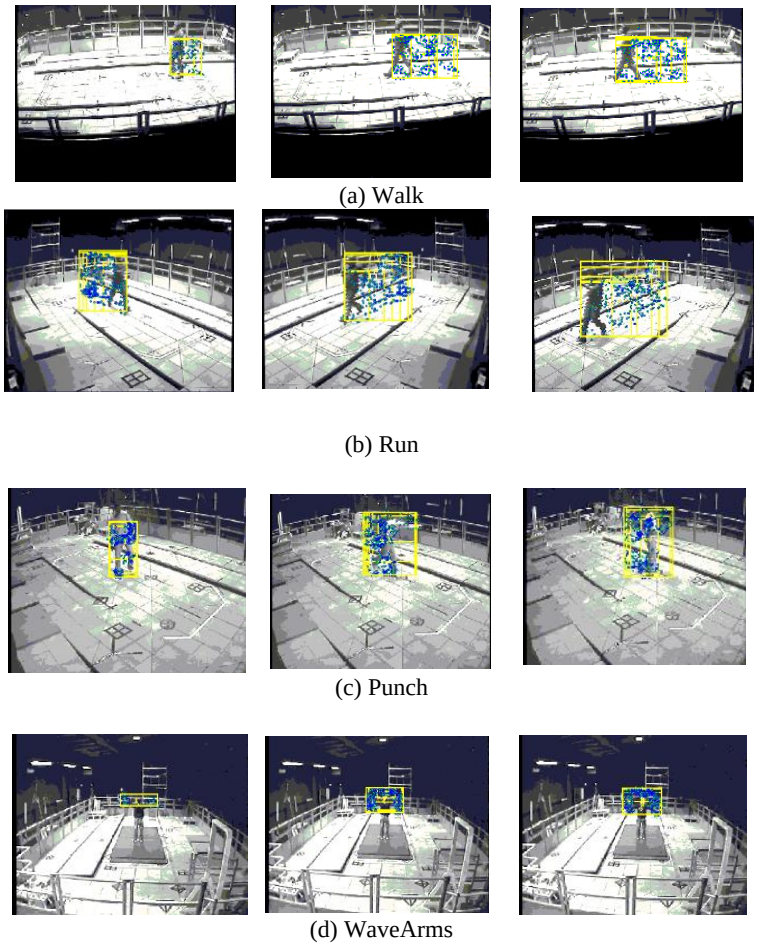
Since, each video frame includes *S* temporal scales. For example, for each frame, there are *S* point clouds of interest. In total, there are 6*S* features from the point clouds of interest. In addition, two other features emanate from the foreground area. As a result, the representation of each frame is 6*S* + 2 features, where *S* is the total number of scales (i.e. 6 features for each scale along with 2 scale-independent features B_t^r and B_t^{Sp}). An overview of the features of the proposed approach are shown in Fig 15.



(a) S = 1

(b) S = 2

Fig. 13: Cloud for different temporal scale S



(a) Walk

(b) Run

(c) Punch

(d) WaveArms

Fig. 14: Examples of the interest points clouds

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

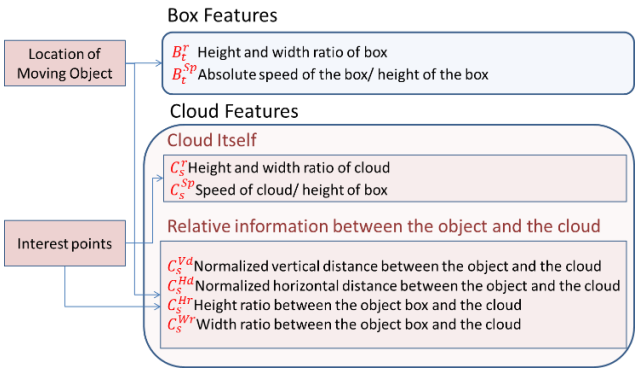


Fig. 15: Overview of the features of the proposed approach

C. *Quantization*

A total $(6S + 2)T$ features are used to represent the whole action sequence, which leads to a very high-dimensional feature space. The high dimension feature space can be caused by over fitting and leads to poor recognition performance. If $S = 6$, we observe one of all the features in all the datasets separately using the empirical cumulative distribution function [39], as shown in Fig 16. The empirical cumulative distribution function reduces the feature space dimension, and more importantly, makes the system representation less sensitive to feature noises and invariant to duration T for each action sequence. In particular, the proposed system separates the empirical cumulative distribution function into N_b portions.

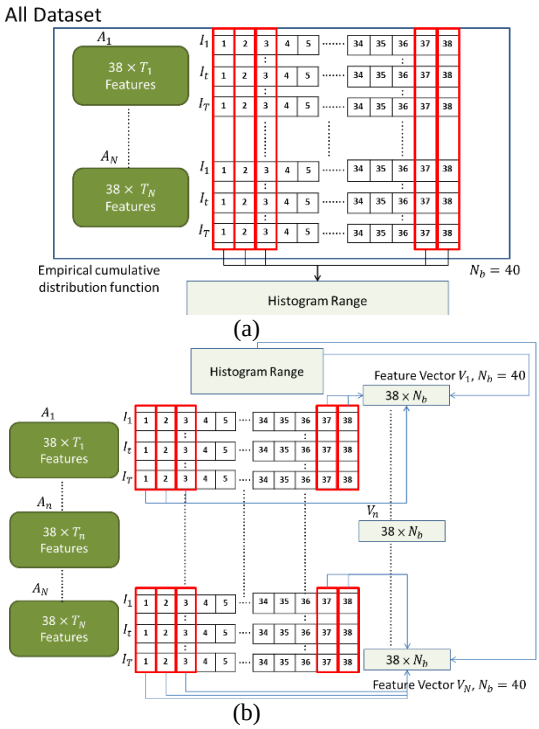


Fig. 16: Overview of quantization (a) A histogram range is produced by observing one of all features in all the dataset separately. (b) Each action sequence A is represented as $(6S+2)$ N_b features.

VI. FEATURE REDUCTION AND CLASSIFICATION

In offline training, the proposed system stores quantized feature vectors, as described in Section IV.C. In online testing, the proposed system uses the histogram range of the training database and transforms the testing data A_{test} to a feature vector V_{test}. Three classifiers are separately used to recognize the testing data for different recognition rates. Fig 17 shows an overview of the feature reduction and classification. The three classifiers are the nearest neighbor classifier (NNC), the Gaussian mixture model classifier (GMMC) and the nearest mean classifier (NMC). This section discusses the different classifiers.

A. *Nearest Neighbor Classifier*

NNC is used widely for action recognition by computing the absolute distance between the testing vector and all of the training vectors. Majority voting is used to classify the object, and usually, the object is classified to the class which was voted the most common amongst its k nearest neighbors. Fig 18 shows an overview using NNC to obtain the most similar action to the testing film. In particular, set $K = 5$ for WEIZMANN dataset, $K = 3$ for KTH dataset and $K = 6$ for MuHAVi dataset. However, it takes a long time at the recognition stage using NNC if there are a large number of training samples because NNC needs to compare whole feature vectors in the database.

B. *Gaussian Mixture Model Classifier*

To reduce the quantity of the feature vectors another method is to use GMMC to model the training data to speed up the recognition time and to utilize *k* Gaussian functions to model each feature of the feature vectors in the database. The result is obtained using the maximum probability value which is summed up by the probability values of each feature, as shown in Fig 19. In particular, three Gaussian functions are set for the KTH dataset, three Gaussian functions for the WEIZMANN dataset and four Gaussian functions for the MuHAvi dataset.

C. *Nearest Mean Classifier*

Another method, the NMC, uses minimum distance between the testing vector and training vectors which is the mean value of the feature vectors of the same action and the same view. An absolute distance is chosen for the recognition decision, as shown in Fig 20. Therefore, NMC is more suitable for the proposed system for real-time recognition and has a better recognition rate. Moreover, the dimension of the subject is reduced to one, which improves performance and results in more efficient recognition.

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

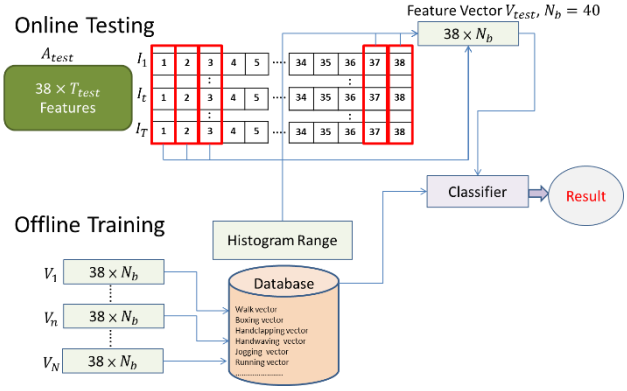


Fig. 17: Overview of feature reduction and classification

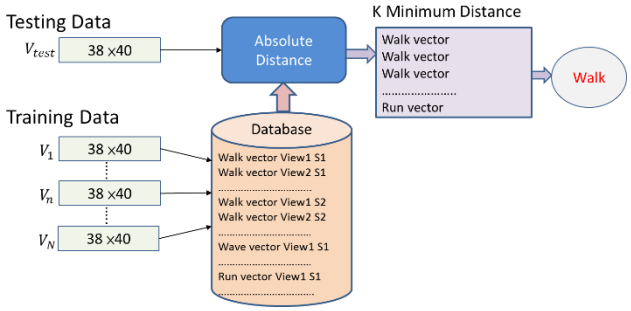


Fig. 18: Overview of NNC for the proposed work

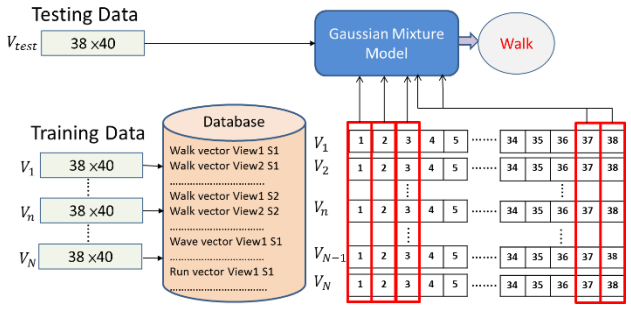


Fig. 19: Overview of GMMC for the proposed work

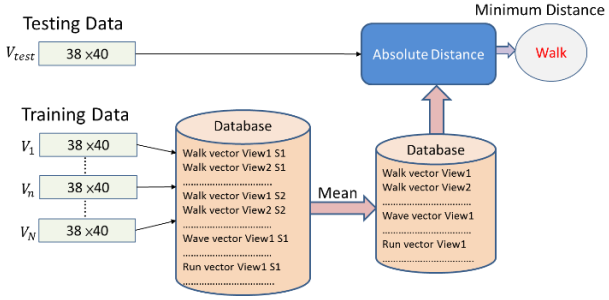


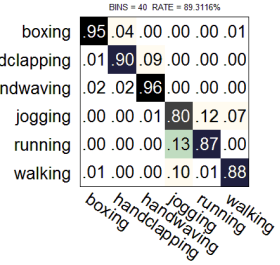
Fig. 20: Overview of NMC for the proposed work

VII. EXPERIMENT RESULT

In this section, several results of action recognition are presented. This section details the recognition rate for subject invariance and for view invariance in section and section VII.B, respectively. The algorithm was implemented on a PC platform with Intel Core i5 3.3GHz and 8GB RAM. The development tool was MATLAB2010 and the operating system was Windows 7. All of the testing inputs are uncompressed AVI video files. The resolution of the video frame is based on the testing datasets. In order to construct multi-scale interest point clouds, *N s* was set to 5 and the total number of scales was 6. This gives 38 features, and a 40-bin histogram can be generated through linear quantization for each feature, for instance, the total features can be represented in 1520 dimensional space.

A. Subject Invariance Evaluation

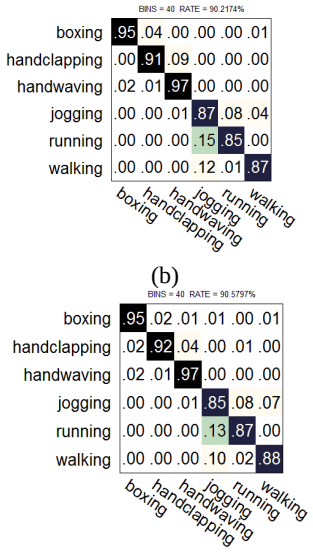
To evaluate subject invariance, the Leave-One-Out Cross-Validation (LOOCV) scheme is adopted to compute the recognition rates. It selects a group of clips from a single subject in a dataset as the testing data, and the rest of the clips are the training data. The repeated progress ensures that each group of clips in the dataset is used once as the testing data. For the KTH dataset, the clips of 24 subjects were used for training and the clips of the remaining subjects were used for validation. For the WEIZMANN dataset, the training set contains 8 subjects. For the MuHAVi dataset, 5 of the 17 actions (Walk- TurnBack, Run-Stop, Punch, CrawlOnKnees, WaveArms) were chosen as the experimental data and the clips of 6 subjects were used for training and the clips belonging to the remaining subjects were used for validation. The results of using NNC, GMMC and NMC for the KTH dataset, WEIZMANN dataset and MuHAVi dataset are shown in Fig 21, Fig 22, Fig 23 and Table I. In particular, NMC obtained a recognition rate of 90.5797% for the KTH dataset, 95.5556% for the WEIZMANN dataset and 97.5% for the MuHAVi dataset. Table II compares the proposed approaches with the existing approaches, the results showing that GMMC and NMC outperform the existing methods on the WEIZMANN and MuHAVi dataset.



(a)

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access



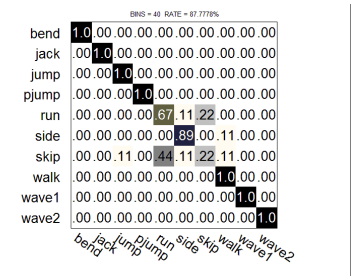
(b)

(c)

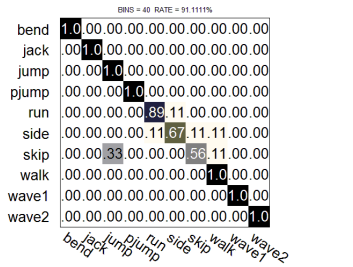
Fig. 21: Recognition performance of the proposed approach for KTH dataset measured using confusion matrices: (a) NNC (b) GMMC (c) NMC

Table I Recognition performance of our approach by using NNC, GMMC, NMC			
	KTH	WEIZMANN	MuHAVi
Our approach (NNC)	89.31%	87.78%	92.50%
Our approach (GMMC)	90.21%	91.11%	93.21%
Our approach (NMC)	90.58%	95.56%	97.50%

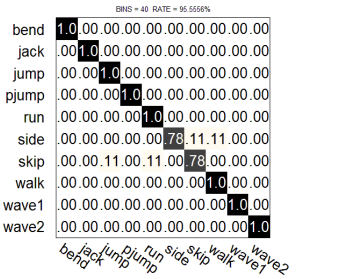
Table II Comparative results on the KTH, WEIZMANN, MuHAVi datasets for subject invariance			
	KTH	WEIZMANN	MuHAVi
Our approach (NMC)	90.58%	95.56%	97.50%
S. Gong et al.[18]	93.17%	96.66%	91.78%
Niebles et al. [15]	83.30%	90.00%	--
Dollar et al. [14]	81.17%	85.20%	--
Zhang et al. [23]	91.33%	92.89%	--
Gilbert et al. [21]	89.92%	--	--
Savarese et al. [22]	86.83%	--	--
Nowozin et al.[24]	84.72%	--	--



(a)



(b)



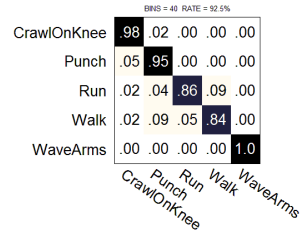
(c)

Fig. 22: Recognition performance of the proposed approach for WEIZMANN dataset measured using confusion matrices: (a) NNC (b) GMMC (c) NMC

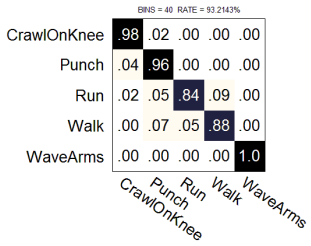
Table III Recognition performance of each view in MuHAVi dataset using GMMC and NMC		
Training View	GMMC	NMC
View1	68.57%	70.36%
View2	68.93%	75.00%
View3	80.00%	83.21%
View4	82.86%	83.57%
View5	77.14%	82.86%
View6	80.71%	85.00%
View7	80.36%	80.36%
View8	80.36%	82.14%

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

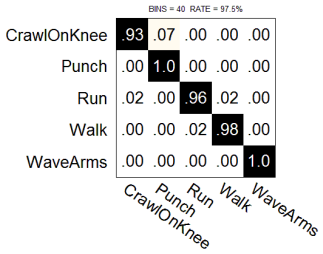
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access



(a)



(b)



(c)

Fig. 23: Recognition performance of the proposed approach for MuHAVi dataset measured using confusion matrices: (a) NNC (b) GMMC (c) NMC

Table IV Comparative results on the MuHAVi dataset for view invariance evaluation	
	MuHAVi
Our approach (NMC)	81.43%
S. Gong et al.[18]	72.85%
A. Eweiwi et al. [25]	77.50%

B. View Invariance Evaluation

To evaluate the proposed method in relation to view invariance, a group of clips from a single view in a dataset is employed as the training data and the remaining clips are the frames, of each action as the testing data. This was repeated so that each group of clips in this dataset is used once as the training data. Five actions out of 17 in the MuHAVi dataset were chosen as the experimental data similar to the subject invariance evaluation. Then, one of the eight views in the MuHAVi dataset is utilized intraining and the other view is utilized in testing. This procedure is repeated for all 8 views and the resulting recognition rates are then averaged. The recognition rates are 78.2143% and 81.4286% using GMMC and NMC, respectively (as shown in Fig 24). Table III shows the recognition rate of each view using GMMC and NMC. The recognition rates of training view3, view5, view6 and view8 are better than the others. These views contain more information than the other four views which allows them to be more robust to view change. Table IV compares the results with the existing approaches. It can be seen that the proposed method is better than the others.

We also evaluate the proposed approach in terms of its robustness against different cameras and evaluate it in terms of view invariance using cross dataset testing. There are three similar actions (Walk, Run, Wave) in the three datasets, including different scenes as previously discussed. The results shown in Table V and Table VI indicate that the recognition rate of training the MuHAVi dataset is better than training the KTH and WEIZMANN datasets since the MuHAVi dataset contains many views. However, the recognition rate of the testing MuHAVi dataset is worse than the testing KTH and WEIZMANN datasets since the MuHAVi dataset tests many actions belonging to different views which are not included in the KTH and WEIZMANN datasets.

Table V Recognition rate of the proposed approach for cross dataset testing using GMMC			
Train / Test	KTH	WEIZMANN	MuHAVi
KTH	90.22%	97.00%	93.33%
WEIZMANN	85.19%	91.11%	96.30%
MuHAVi	83.14%	84.50%	93.21%

Table VI Recognition rate of the proposed approach for cross dataset testing using NMC			
Train / Test	KTH	WEIZMANN	MuHAVi
KTH	90.58%	97.00%	96.30%
WEIZMANN	84.67%	95.56%	95.56%
MuHAVi	84.53%	88.50%	97.50%

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

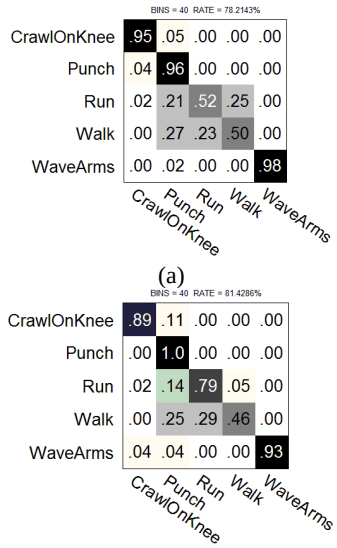


Fig. 24: Recognition performance of view invariance evaluation using confusion matrices: (a) MuHAvi dataset using GMMC (b) MuHAvi dataset using NMC

C. *Auto Labeling*

The proposed method utilizes a mechanism which can watch a person's actions in an image sequence and separate these actions automatically. Firstly, it adopts four different temporal scales ((1/4)T-frames, (1/2)T-frames and T-frames) of each action to be the training data for the offline training. Secondly, feature vectors of the four temporal scales belonging to each action are produced using the function described in Section IV and Section V. Then, these feature vectors are placed into different temporal scale databases.

For the online testing, first, the system scans the image sequence using a scanning window whose temporal scale is (1/4) T -frames. Second, the (1/4)T -frames window $W_{(1/4)T}$ is transformed to a feature vector using the function detailed in Section IV and Section V. Then, in the classification stage, NMC is used to classify the feature vectors from the T-frames database. Actions are classified as candidate actions if similarity S is over 70%. Similarity S is defined as:

$$S = \frac{F \times N_b - D}{F \times N_b} \qquad (15)$$

where F is the number of features, N_b is the number of bins and D is the absolute distance between the testing feature vector and the training feature vector. However, if similarity S is below 70% the (1/4) T-frames scanning window $W_{(1/4)T}$ skips I frames to find other actions from the other images.

In the experiment, set F = 38, N_b = 40 and I = 15. As soon as some actions produced by the (1/4)T -frames scanning window are deemed to be candidate actions, the system uses (1/2)T -frames scanning window $W_{(1/2)T}$ to scan the next (1/4)T-frames and the previous (1/4) T -frames. In the classification stage, NMC is utilized to classify the feature vector from (1/2)T -frames database of the candidate actions. Similar to the (1/4) T–frames scanning window $W_{(1/4)T}$, the candidate actions remain candidate actions if similarity S is over 70%. Then, set beginning of the action from testing image sequences using the first index of (1/2)T -frames scanning window. (3/4)T -frames, T-frames and (5/4) T -frames scanning windows are used to scan the images and classify the produced feature vectors from the T-frames database. The maximum value similarity S is used to obtain the result. The difference rate R is used to find the end of the action which occurs when the difference R is over 10%. The difference rate R is defined as

$$R = \frac{D_c - D_l}{D_c} \qquad (16)$$

where D_c is the absolute distance of the feature vector between the current scanning window and the training database; D_l is the absolute distance of the feature vector between the last scanning window and the training database. Finally, the system labels one of actions in the image sequences and uses (1/4) T -frames scanning window to find the next action in the video. In the experiment, set T = 100 and Fig 25 shows an overview of auto labeling.

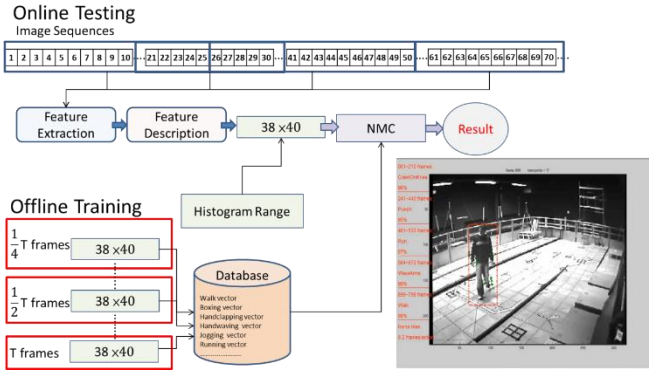


Fig. 25: Overview of auto labeling

VIII. CONCLUSION

This paper presents an approach for real-world applications which automatically labels the beginning and ending of an action sequence. The system uses the proposed view-invariant features to address multi-view action recognition from different perspectives for accurate and robust action recognition. The view-invariant features are obtained by extracting holistic features from different temporal scale clouds, which are modeled on the explicit global, spatial and temporal distribution of interest points. The experiments on the KTH and WEIZ- MANN datasets demonstrate that using view-invariant features

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

obtained by extracting holistic features from clouds of interest points is highly discriminative and more robust for recognizing actions under different view changes. The experiments also show the proposed approach performs well with cross-tested datasets using previously trained data, which means there is no need to re-train the system if the scenario changes.

Acknowledgements

This work was supported in part by the Australian Research Council (ARC) under discovery grant DP180100670 and DP180100656; in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0022 and W911NF-10-D-0002/TO 0023; in part by the Taiwan Ministry of Science and Technology under Grant Number: MOST 106-2218-E-009-027-MY3 and MOST 106-2221-E-009-016–MY2.

REFERENCE

[1] K. G. Derpanis, M. Sizintsev, K. J. Cannons, and R. P. Wildes, "Action spotting and recognition based on a spatiotemporal orientation analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 527–540, 2013.

[2] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 883–897, 2011.

[3] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1860–1870, 2013.

[4] Y. Yang, I. Saleemi, and M. Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1635–1648, 2013.

[5] Z. Jiang, Z. Lin, and L. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 3, pp. 533–547, 2012.

[6] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," IEEE Transactions on Image Processing, vol. 22, no. 6, pp. 2479–2494, 2013.

[7] Y. Chen, Z. Li, X. Guo, Y. Zhao, and A. Cai, "A spatio-temporal interest point detector based on vorticity for action recognition," in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6

[8] S. Samanta and B. Chanda, "Space-time facet model for human activity classification," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1525–1535, 2014

[9] Z. Moghaddam and M. Piccardi, "Histogram-based training initialisation of hidden markov models for human action recognition," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE Inter- national Conference on*. IEEE, 2010, pp. 256–261

[10] Y. Wang, L. Wu, and X. Huang, "Action recognition using tri-view constraints," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*. IEEE, 2011, pp. 107–112

[11] M. N. Kumar and D. Madhavi, "Improved discriminative model for view-invariant human action recognition," Intl. Journal of Computer Science & Engineering Technology, vol. 4, pp. 1263–1270, 2013.

[12] F. Zhang, Y. Wang, and Z. Zhang, "View-invariant action recognition in surveillance videos," in *Pattern Recognition (ACPR), 2011 First Asian Conference on*. IEEE, 2011, pp. 580–583

[13] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2012

[14] N. Ikizler-Cinbis and S. Sclaroff, "Web-based classifiers for human action recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1031–1045, 2012

[15] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 236–243, 2013

[16] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 32–36

[17] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007

[18] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*. IEEE, 2010, pp. 48–55

[19] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *null*. IEEE, 2003, p. 726

[20] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8

[21] C. Rao and M. Shah, "View-invariance in action recognition," in *Com- puter Vision and Pattern*

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2809552, IEEE Access

Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 2. IEEE, 2001, pp. II–II

[22] A. Ali and J. Aggarwal, "Segmentation and recognition of continuous human activity," in *Detection and recognition of events in video, 2001. Proceedings. IEEE Workshop on*. IEEE, 2001, pp. 28–35

[23] D. Ramanan and D. A. Forsyth, "Automatic annotation of everyday movements," in *Advances in Neural Information Processing Systems*, 2004, pp. 1547–1554

[24] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 144–149

[25] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 166–173

[26] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 984–989

[27] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004

[28] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM International Conference on Multimedia*. ACM, 2007, pp. 357–360

[29] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *Computer Vision–ECCV 2008*, pp. 650–663, 2008

[30] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recog- nition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72

[31] A. Eweiwi, S. Cheema, C. Thurau, and C. Bauckhage, "Temporal key poses for human action recognition," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1310–1317

[32] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8

[33] A. Iosifidis, A. Tefas, and I. Pitas, "Neural representation and learning for multi-view human action recognition," in *the 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012, pp. 1–6

[34] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011

[35] Y. Lu, Y. Li, Y. Shen, F. Ding, X. Wang, J. Hu, and S. Ding, "A human action recognition method based on Tchebichef moment invariants and temporal templates," in *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2012 4th International Conference on*, vol. 2. IEEE, 2012, pp. 76–79

[36] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 2. IEEE, 1999, pp. 246–252

[37] J. R. Parker, *Algorithms for image processing and computer vision*. John Wiley & Sons, 2010

[38] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1948–1955

[39] G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*. SIAM, 2009

[40] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, "Motion context: A new representation for human action recognition," *Computer Vision–ECCV 2008*, pp. 817–829, 2008

[41] A. Gilbert, J. Illingworth, and R. Bowden, "Scale invariant action recognition using compound features mined from dense spatio-temporal corners," in *European Conference on Computer Vision*. Springer, 2008, pp. 222–233

[42] S. Savarese, A. DelPozo, J. C. Niebles, and L. Fei-Fei, "Spatial-temporal correlatons for unsupervised action classification," in *Motion and video Computing, 2008. WMVC 2008. IEEE Workshop on*. IEEE, 2008, pp. 1–8

[43] S. Nowozin, G. Bakir, and K. Tsuda, "Discriminative subsequence mining for action classification," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* IEEE, 2007, pp. 1–8.

2169-3536 (c) 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.