Abstract—In this paper a novel algorithm for human action recognition is presented. This approach is based on Two- dimensional principal component analysis (2DPCA) and vector quantization (VQ) in the spatial-temporal domain. This method reduces computational complexity by a factor of 98, while maintaining the storage requirement and the recognition accuracy, compared with some of the most recent approaches in the field. Experimental results applied on the Weizmann dataset confirm the excellent properties of the proposed algorithm.

I. INTRODUCTION

Human action recognition is one of the most challenging research problems in the field of computer vision. This research area has many applications, such as video indexing, video surveillance, human-computer interface design, gesture recognition, analysis of sports videos, and the development of intelligent environments.

A review on action recognition research area has been presented in a number of survey papers ([1], [2], [3]). In general, approaches for human action analysis can be categorized into four main approaches. A brief summary of these methods will be presented in the following paragraphs.

In 1992, Yamato et al [4] introduced appearance-based approaches for Human action recognition. In this technique learning the appearance model of the human body and matching it explicitly to images in a target video sequence for action recognition. Recently a new approach was presented by Dhillon et al [5]. This method combines information from appearance-based approach and motion of human body parts. It tracks the human body parts by using mixture particle filters and then clustering the particles using local non-parametric clustering. This model used to build one histogram per video sequence which provides the characteristic information to discriminate among various human actions.

Optical Flow-based approaches and motion information for action recognition have been used in ([6], [7], [8], [9], [10], [11]). In 2003 Efros et al [6] introduced a motion descriptor based on optical flow measurements in a spatio-temporal volume for each stabilized human figure, and an associated similarity measure used in a nearest-neighbor framework. A recent approach [12] introduces a set of kinematic features that are derived from the optical flow, for human action recognition in videos. Each kinematic feature, computed from the optical flow of a sequence of images, and gives rise to a spatio-temporal pattern kinematic modes can be computed by performing Principal Component Analysis (PCA) on the spatio-temporal volumes of the kinematic features. A nearest neighbor algorithm is used for classification, using kinematic mode-based feature space and the coordinates of the video in that space for classification.

Temporal-based approaches depend on periodicity analysis of human actions, [13] and thus limited to cyclic actions. Other approaches using shape-based representations for action recognition depends on key frames or eigenshapes of foreground silhouettes ([14], [15]), where action consists of a series of poses that are detectable from a single frame. Each pose can be encoded separately using the shape features, and single frame recognition can be extended to multiple frames for robust action recognition.

Further modification lead to, a spatio-temporal representations, this approaches based on video sequences a human action generates a space-time shape in the space-time volume ([16], [17], [18], [19]). Such as in ([16], [17]) these shapes are induced by a concatenation of 2D silhouettes in the space-time volume and contain both the spatial information about the pose of the human figure at any time, as well as the dynamic information. The limitation of these methods is that the inability to detect the intra-class variability of actions. Recently Mikel et al. [20] attempt to make general capturing intra-class variability by synthesizing a single action filter for a given action class using Maximum Average Correlation Height (MACH) filter. Analyzing using the Clifford Fourier transform in the frequency domain, and avoid the high computational cost commonly incurred in template-based approaches.

In 2004 Yang et al [21] proposed the two dimensional PCA (2DPCA) technique for facial recognition, which has many advantages over the PCA method. It is simpler for image feature extraction, better in recognition rate and more efficient in computation. However, it is not as efficient as PCA in terms of storage requirements.

In a previous contribution by Abdelwahab and Mikhail [22], a Transform Domain 2DPCA (TD2DPCA) employing VQ for facial image recognition was presented. TD2DPCA algorithm reduced the storage requirements by 90%

Human Action Recognition Employing 2DPCA and VQ in the Spatio-Temporal Domain

Mohamed A. Naïel(1)  Moataz M. Abdelwahab(1)  Wasfy B. Mikhael(2)

(1)School of Communications and Information Technology Nile University 6th October, Egypt
mohamed.naiel@nileu.edu.eg, mabdelwahab@nileuniversity.edu.eg,
mikael@mail.ucf.edu

(2)School of Electrical Engineering and Computer Science University of Central Florida Orlando, FL, USA

Abstract—In this paper a novel algorithm for human action recognition is presented. This approach is based on Two-Dimensional principal component analysis (2DPCA) and vector quantization (VQ) in the spatial-temporal domain. This method reduces computational complexity by a factor of 98, while maintaining the storage requirement and the recognition accuracy, compared with some of the most recent approaches in the field. Experimental results applied on the Weizmann dataset confirm the excellent properties of the proposed algorithm.

I. INTRODUCTION

Human action recognition is one of the most challenging research problems in the field of computer vision. This research area has many applications, such as video indexing, video surveillance, human-computer interface design, gesture recognition, analysis of sports videos, and the development of intelligent environments.

A review on action recognition research area has been presented in a number of survey papers ([1], [2], [3]). In general, approaches for human action analysis can be categorized into four main approaches. A brief summary of these methods will be presented in the following paragraphs.

In 1992, Yamato et al [4] introduced appearance-based approaches for Human action recognition. In this technique learning the appearance model of the human body and matching it explicitly to images in a target video sequence for action recognition. Recently a new approach was presented by Dhillon et al [5]. This method combines information from appearance-based approach and motion of human body parts. It tracks the human body parts by using mixture particle filters and then clustering the particles using local non-parametric clustering. This model used to build one histogram per video sequence which provides the characteristic information to discriminate among various human actions.

Optical Flow-based approaches and motion information for action recognition have been used in ([6], [7], [8], [9], [10], [11]). In 2003 Efros et al [6] introduced a motion descriptor based on optical flow measurements in a spatio-temporal volume for each stabilized human figure, and an associated similarity measure used in a nearest-neighbor framework. A recent approach [12] introduces a set of kinematic features that are derived from the optical flow, for human action recognition in videos. Each kinematic feature, computed from the optical flow of a sequence of images, and gives rise to a spatio-temporal pattern kinematic modes can be computed by performing Principal Component Analysis (PCA) on the spatio-temporal volumes of the kinematic features. A nearest neighbor algorithm is used for classification, using kinematic mode-based feature space and the coordinates of the video in that space for classification.

Temporal-based approaches depend on periodicity analysis of human actions, [13] and thus limited to cyclic actions. Other approaches using shape-based representations for action recognition depends on key frames or eigenshapes of foreground silhouettes ([14], [15]), where action consists of a series of poses that are detectable from a single frame. Each pose can be encoded separately using the shape features, and single frame recognition can be extended to multiple frames for robust action recognition.

Further modification lead to, a spatio-temporal representations, this approaches based on video sequences a human action generates a space-time shape in the space-time volume ([16], [17], [18], [19]). Such as in ([16], [17]) these shapes are induced by a concatenation of 2D silhouettes in the space-time volume and contain both the spatial information about the pose of the human figure at any time, as well as the dynamic information. The limitation of these methods is that the inability to detect the intra-class variability of actions. Recently Mikel et al. [20] attempt to make general capturing intra-class variability by synthesizing a single action filter for a given action class using Maximum Average Correlation Height (MACH) filter. Analyzing using the Clifford Fourier transform in the frequency domain, and avoid the high computational cost commonly incurred in template-based approaches.

In 2004 Yang et al [21] proposed the two dimensional PCA (2DPCA) technique for facial recognition, which has many advantages over the PCA method. It is simpler for image feature extraction, better in recognition rate and more efficient in computation. However, it is not as efficient as PCA in terms of storage requirements.

In a previous contribution by Abdelwahab and Mikhail [22], a Transform Domain 2DPCA (TD2DPCA) employing VQ for facial image recognition was presented. TD2DPCA algorithm reduced the storage requirements by 90%
compared to the 2DPCA method [21]. Also, the TD2DPCA improved the computational speed by 50% while maintaining the excellent recognition accuracy. In this paper we applied the algorithm in the spatial domain and modified this algorithm to be applied on videos for human action recognition.

In this paper, a novel spatio-temporal based approach for human action recognition employing principal component analysis and vector quantization is introduced. This technique reduces computational complexity to a great deal, while maintaining the storage requirement and the recognition accuracy compared to most recent approaches in the field. A set of experiments performed on publicly available dataset, confirm the excellent properties of the algorithm.

This paper is organized as follows: Section II discusses the proposed algorithm. Section III shows experimental results and analysis obtained by testing the new algorithm on the Weizmann dataset. The conclusions are presented in section IV.

II. THE PROPOSED ALGORITHM

In this method human action is recognized using 2DPCA and Vector Quantization, performed directly on the video silhouette, obtained from foreground segmentation. The algorithm is divided into two modes, the training mode and the testing mode, as follows:

A. Training Mode

In the training mode, the training videos representing different actions are introduced to the system. Matrix $M$, denotes the input silhouette spatio-temporal for all the training sequences.

The features of the data base are extracted, grouped, and stored as described following the steps below.

1) Read the input videos in matrix $M$ of size $(m \times n \times k)$, where $m$ represents the number of rows for every sequence, $n$ the number of columns for every sequence, $k$ the size of the sequence (number of frames for all the training videos).

2) The covariance matrix $S$, of size $(n \times n)$, for the $k$ training frames is calculated as follows.

$$ S = \frac{1}{k} \sum_{j=1}^{k} (M_{mxnxj} - \overline{A})^T (M_{mxnxj} - \overline{A}) $$

(1)

Where $\overline{A}$ is the mean matrix, of all the $k$ training sequences, of size $(m \times n)$.

3) A set of $r$ eigenvectors, $V_q$, of size $(1 \times n)$ corresponding to the dominant eigenvalues $\lambda_q$, where $q=\{1, 2, ..., r\}$, is obtained for matrix $S$.

4) Store the matrix $V = [V_1, V_2, ..., V_r]$.

5) The matrix $N$, with dimensions $(m \times n \times h)$, representing the $i^{th}$ training video is formed. Where $i=\{1, 2, ..., B\}$, $B$ the maximum number of training videos and $h$ represents the number of frames in the $i^{th}$ video.

6) Project every frame in matrix $N$, denoted by $G$, of size $(m \times n)$, where $y=\{1, 2, ..., h\}$, on the matrix $V$ to obtain the feature matrix $F_y$, of size $(m \times r)$.

$$ F_y = G V $$

(2)

7) Every feature matrix $F_y$ is concatenated to produce feature vector, $f$, the size of this vector is $(1 \times p)$, where $p=mr$.

8) The feature matrix, $R$, for a given video sequence, of size $(h \times p)$, is constructed, where each row consists of the feature vector $f$, where $y=\{1, 2, ..., h\}$.

$$ R = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_h \end{bmatrix} $$

(3)

9) Employing VQ, a centroid $C_{i}=[x^{(i)}_1, x^{(i)}_2, ..., x^{(i)}_p]$, of size $(1 \times p)$ representing the feature matrix $R$ is obtained.

10) Repeat steps 5 to 9 for every action.

11) Store the centroids, $C_{i}$, $i=\{1, 2, ..., B\}$, representing each video sequence.

B. Testing Mode

In the testing mode the input video is tested according to the following steps:

1) Read the input video in matrix $N_{i}$, of size $(m \times n \times h)$, where $N_{i}$ represents the input silhouette for one testing sequence.

2) Repeat steps from 5 to 9 in the training mode, to obtain $C_{i}$, where $C_{i}$ represents the centroid of the input action after projection of $N_{i}$ on $V$.

3) A nearest neighbor classifier is used to measure the distance between the resulted centroid, $C_{i}$ and the stored centroids, $C_{j}$, $i=\{1, 2, ..., B\}$, using the Euclidean distance as follows:

$$ D_{i}(C_{i}, C_{j}) = \sum_{t=1}^{p} \left| x^{(t)}_{i} - x^{(t)}_{j} \right|^2 $$

(4)

Where $\left| x^{(t)}_{i} - x^{(t)}_{j} \right|$ denotes the Euclidean distance between the two elements $x^{(i)}_{t}$ and $x^{(j)}_{t}$.

4) The minimum distance $D_{i}$ corresponds to the estimated action of the $i^{th}$ video.

III. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed Algorithm was applied to the Weizmann Action Dataset [16]. Two experiments (I, II) were conducted. Experimental results were compared to methods that were recently published.

The Weizmann dataset consists of 90 low-resolution (180 x 144, 50 fps) video sequences showing nine different actors, each performing 10 natural actions such as walk, run, jump forward (jump), gallop sideways, bend, wave one hand (wave1), wave two hands (wave2), jump in place, jump-jack, and skip, as shown in Fig. 1.
Two experiments, I and II, were conducted on the available aligned dataset [23] shown in Fig. 2. This aligned dataset consists of 90 low-resolution (120 x 90, 50 fps). The available silhouettes contained “leaks” and “intrusions” due to imperfect subtraction, shadows, and color similarities with the background.

In experiment I, we divided all actions into 2-fold cross validation, by selecting four videos per action for training, and the remaining five videos for testing. In our training algorithm, \( r=20, C_i \), where \( i=[1, 2, ..., 40] \) each of size \((1 x 2400)\), and matrix \( V \) of size \((90 x 20)\).

In the testing mode, the remaining video was used for testing the proposed algorithm. A sliding window defined by 28 frames form one space-time cube, \( N_s \), with zero overlap between consecutive space-time cubes. This tests the algorithm performance on short sequences. The confusion matrix, in Table I(a), shows that the average recognition accuracy is 95.42%. It can be used to classify real-time scenarios. The average testing time was 18.8 milliseconds per cube. This excellent performance is obtained while having a minimum storage requirement, 40 centroids each of size \((1 x 2400)\). This approach shows that eight actions are 100% accuracy.

In experiment II, a leave-one-out cross validation technique is employed. Where the tested video was removed from the training videos. In our training algorithm, \( r=20, C_i \), where \( i=[1, 2, ..., 89] \) each of size \((1 x 2400)\), and matrix \( V \) of size \((90 x 20)\).

In the testing mode, the remaining video was used for testing the proposed algorithm. A sliding window defined by 28 frames form one space-time cube, \( N_s \), with zero overlap between consecutive space-time cubes. The confusion matrix, in Table I(b), shows that the average recognition accuracy is 96.19%. The average testing time was 18.8 milliseconds per cube, which is faster than the other techniques by 98 times [16], [20]. This excellent performance is obtained while having a minimum storage requirement, 89 centroids each of size \((1 x 2400)\). This approach shows that seven actions are 100% accuracy. Also wave one hand and wave two hands have high similarity.

Table II shows that we are maintaining the excellent recognition accuracy compared with other recent published approaches. This is achieved with minimum amount of training videos.

In general, table III shows that we are faster than the other available techniques. The average running time of the tested, on Pentium 4, 3.0 GHz, consistent with the reports published [20], [16], and [17], was 186.9 milliseconds which is 98 times faster than the best available record.
TABLE II. COMPARISON OF THE BEST OVERALL ACCURACY AND TESTING TECHNIQUE ON WEIZMANN DATASET WITH PREVIOUS WORK

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Testing technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed algorithm</td>
<td>95.42%</td>
<td>40-50 Split</td>
</tr>
<tr>
<td>Dhillon et al [5]</td>
<td>88.50%</td>
<td>63-27 Split</td>
</tr>
<tr>
<td>Proposed algorithm</td>
<td>96.19%</td>
<td>Leave one out</td>
</tr>
</tbody>
</table>

| Blank et al [16]     | 97.83%        | Leave one out     |
| Niebles & Fei-Fei [24]| 72.8%        | Leave one out     |
| Yuan et al. [25]     | 92.22%        | Leave one out     |
| Saad & Shah [12]     | 95.75%        | Leave one out     |
| Yang et al [26]      | 92.8%         | Leave one out     |

TABLE III. COMPARISON OF THE AVERAGE TESTING RUNTIME ON WEIZMANN DATASET WITH PREVIOUS WORK

<table>
<thead>
<tr>
<th>Method</th>
<th>Average testing runtime</th>
<th>Video size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>186.9 milliseconds</td>
<td>144 x 180 x 200</td>
</tr>
<tr>
<td>Mikel et al</td>
<td>18.65 seconds</td>
<td>144 x 180 x 200</td>
</tr>
<tr>
<td>Blank et al</td>
<td>30 seconds</td>
<td>110 x 70 x 50</td>
</tr>
<tr>
<td>Blank et al</td>
<td>30 minutes</td>
<td>144 x 180 x 200</td>
</tr>
</tbody>
</table>

IV. CONCLUSIONS

A novel algorithm for human action recognition is presented. Our approach is based on 2DPCA and VQ applied in spatial-temporal domain. This method reduced the computational complexity by a factor of 98, while maintaining the recognition accuracy and storage requirements, compared with the most recent approaches. Experimental results performed on the Weizmann dataset confirm the excellent properties of the proposed algorithm showing that this method is robust and work with small number of training sequences. This dramatic reduction in the computational complexity for the tested videos promote this algorithm for real-time applications. For future work, our proposed method can be applied using multi-transform domains, where multi-criteria can be extracted to improve the recognition accuracy.

REFERENCES