ACTION RECOGNITION AND IDENTITY USING CHEBYSHEV MOMENTS

Khawlah Hussein Alhamzah 1,2, Tianjiang Wang 2

1 Dept of Computer Science, University of Basrah, Iraq
2 School of Computer Science, Huazhong University of Science and Technology (HUST), China

ABSTRACT:
The goal of this paper is to build a novel framework of human action recognition and identity at the same time for surveillance video applications. Since, most researches focus on action recognition algorithms only, without know who makes the action, because it is very difficult to recognize the human face from surveillance camera due to the far distance and no contact with biometrics of humane like face, fingerprint. There is little research found that they separated the two problems, where at the first stage recognize the action and the second stage had to be assigned the identity to the action through a comparison with a previous database stored. We see that this is requires additional computation corresponding to large database and low accuracy. In this paper, we have trained human with his action at the same time, we do this by embedding the watermark to the video of a binary image as 2-D wavelet transform represents human’s identity. Motion energy image (MEI) is used to characterize the motion regions of the video frames, then we proposed to use Chebyshev moments to extract feature vectors of MEI for human actions, which is based on discrete orthogonal properties, that is very importing in recognition accuracy due to mutual independence features with no redundancy information and robust to translation, scaling, rotation and noise. The only post processing is extracting the watermark by inverse discrete wavelet transform (IDWT) to extract the human identity. We used multi-class linear SVM as classifier and KTH dataset.
The action classification in 95% and identity classification is 98.7%.

Key Words: Action recognition, Video watermarking, Discrete Wavelet Transform, Motion Energy Image (MEI), Chebyshev moments, MSE error, SVM.

[1] INTRODUCTION

In recent years has increased the need for applications to human activity, such as surveillance cameras especially in security area, entertainment and healthcare system. Most researches focus on activity recognition algorithms only, without know who makes the activity. The majority of most researches have learned features for accurate representation of human action in the video so that it can represent the action of the human as feature vectors and then feed to the classifier environment [17]. In [2] Khawlah at el., produced a framework to trained action and identity at same time, they used watermark to find the human identity and SIFT as feature descriptors, the feature vector is high dimensionality. In [3] Zhou at el. present a method based on affine SIFT to capture motion at first step then assign identification at second step, the accuracy of action recognition is 81.2% and identity recognition is 58.8%. Moments and
functions of moments have been used as pattern features in a number of applications to meet invariant recognition of two-dimensional image patterns [8, 9]. Hu [29] first introduced moment invariants in 1961, based on methods of algebraic invariants. However, the recovery of the image from these moments is deemed to be quite difficult. Teague [26] has suggested the idea of orthogonal moments to recover the image from moments, and has introduced Zernike moments. In this paper, we used moment invariants as feature descriptors. Given the significant literature in the area, we focusing on the most relevant works of represent actions with moment [8, 9]. In [1] Khawlah at el. produced deep learning for feature learning and watermark for human identity. In [28] Okay at el. used 3D Zernike moments as feature descriptors, they get 3D volumetric data (voxel data), and used the surface information contained in the dataset. The main drawback of space-time descriptors is high dimensionality and efficiency, the amount of preprocessing before moment calculations is: First, the centered of the 3D image is calculated and becomes the origin. The unit ball, in which all moment calculations are carried out, is then defined as a sphere of one meter radius having the same origin. We believe that their model takes many calculations. In [27] T. Wang at el used 2D Zernike moments as feature descriptors for Chinese characters. In [4] Yanan Lu at el. produced a method based on Chebyshev moments and two models of temporal templates; MEI and MHI for feature extraction, the accuracy of action recognition is 93% . In [27] their model, four lower orders of Zernike moments with the highest variance values is utilized to form a feature vector to recognize Chinese characters. In [22] Flusser at el. produced system that used Hu invariant moments to identify aircraft by moment invariants. The Hu moments [29] are more redundancy information in higher order of moments which are not suitable for recognition [22]. In [25] their model used geometric Hu moments to drive invariant moments for gray scale images. In this paper, we used Chebyshev moments [6] as a feature vector and its discrete manner due to quantization error of orthogonal continuous moments like Zernike and the need for a transformation from the image coordinate space to a normalized coordinate space in continuous moments. The orthogonal moments are obtained by weighted averaging of image pixels and thus be expected to be robust against noise. Nowadays, Chebyshev moments are becoming popular for shape recognition because of its discrete property that is suitable to image processing, the only technique that may be compared to moments for binary shape characteristics is Fourier descriptors, which is based on object boundary rather than silhouette (entire contents), may be shown to be more sensitive to boundary variations. The motivation of this work is purely based on the application of Chybeshev moments as feature descriptors recognition. As a result, they are making high fidelity reconstruction (which is a measure of the quality of representation that a certain level of availability of the moment) when used to discrete moments. In this paper, six lower orders of Chebyshev moments will be selected as the recognition features. In this paper we interesting of using moments as feature descriptors of actions because of its highly accurate in recognition and low dimension of feature descriptor without using principal component analysis (PCA) or other technique to reduction the dimension The accuracy of Chebyshev moments as descriptors is assessed by means of image reconstruction. By inspecting the image reconstructed from its set of moments, one can determine the number of moments required to capture the essential characteristics of the image, which characterize the global features that capability of Chebyshev moments. The results of these experiments show that Chebyshev moments give positive improvements and have added advantage over the other moments in consideration of image.

In this work, we trained the action and identity at the same time by a simple method. We find the identity of human as watermark embedded as 2-D wavelet transform in the input video in the training data. Watermarking in the DWT domain [12] can be split into the two ways: embedding of the watermark and extraction of the watermark, the whole framework of our propose algorithm is shown in Figure 1. The basic idea of discrete wavelet transform (DWT) in image process is to multi-differentiated decompose the image into sub-image of different spatial domain and independent frequency district. After the original image has been DWT transformed, it is decomposed into 4 frequency districts which is one low-frequency district(LL) and three high-frequency districts(LH,HL,HH), more details can be found in [12]. We insert the identity image watermark in diagonal sub band because other three parts of DWT like approximations, vertical and horizontal are not suitable to embed a watermark. The quality of the video frames is the same as original frames as we show in Figure 2. Second, we segmented the video frames into binary foreground frames to detect the human by using mixture Gaussian model [24]. To represent where motion the image is moving, we used a motion-energy image (MEI) [13] which describes the spatial distribution of motion energy for a given view of a given action, which is mean a cumulative motion image starting from frame 0 to corresponding frame as shown in Figure 4 for KTH dataset. Shape of the region extracted by MEI characterizes both the occurrence of the action and the viewing angle, which is computed as follows:

\[ E_\tau (x,y,t) = \bigcup_{i=0}^{\tau-1} D(x,y,t-i) \]  

Where D is the binary foreground silhouette difference image, \( \tau \) is duration the temporal extent of an action.

![Figure 1. Framework of proposed algorithm](image1)

![Figure 2. Frames with watermark embedding as 2-D wavelet in the KTH dataset (same as original frames)](image2)
[2.1] CHEBYSHEV MOMENTS

In [29] although Hu first proposed seven moment invariants named Hu moment invariants which have translation, rotation and scale invariance, but Hu moment is not an orthogonal moment, so it has high redundancy of information. we propose a novel feature description that is derived from invariant moments, Chebyshev moments, which capture the underlying shape of the action well, even in the presence of significant distortions and data uncertainty, and robust to translation, rotation and scaling. The two-dimensional Chebyshev moments of order p with repetition q of image intensity function [6, 7, 10, 14, 15]:

\[
\begin{align*}
  f(x, y) \\
  &= \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} T_{pq} \cdot t_p(x) \cdot t_q(y)
\end{align*}
\]

(2)

where the coefficient \( T_{pq} \) are the Chebyshev moments which satisfy the orthogonal property, with

\[
\begin{align*}
  P(n, N) &= \frac{N(N^2 - 1)(N^2 - 2^2) \ldots \ldots (N^2 - n^2)}{2n + 1}, n = 0, \ldots, N - 1
\end{align*}
\]

(3)

To make suitable for image analysis:

\[
\begin{align*}
  T_{pq} &= \frac{1}{\rho(p, N) \cdot \rho(q, N)} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} t_p(x) \cdot t_q(y) \cdot f(x, y) \\
  &= 0, 1, \ldots, N - 1
\end{align*}
\]

(4)

Where

\[
\begin{align*}
  \rho(p, N) &= \frac{N \left(1 - \frac{1^2}{N^2}\right) \left(1 - \frac{2^2}{N^2}\right) \ldots \ldots \left(1 - \frac{p^2}{N^2}\right)}{2p + 1}
\end{align*}
\]

(5)

And the scaled Chebyshev polynomials \( t_p(x) \) are computed using the following recurrence relation:

\[
\begin{align*}
  t_0(x) &= 1 \\
  t_1(x) &= \frac{2x + 1 - N}{N} \\
  t_p(x) &= \frac{(2p-1)t_1(x)t_{p-1}(x) - (p-1)\left[1 - \left(\frac{(p-1)^2}{N^2}\right)\right]t_{p-2}(x)}{p}, p > 1
\end{align*}
\]

(6)

(7)

(8)
We use only one temporal template that is MEI for extract action features, that is enough to recognize actions, because we don’t need to use another temporal template; MHI (motion history image), which is decide how action is happened. The usefulness of these moments in our application is that they are used to pre-process images in order to make their features invariant to scale and translation transformations. The orthogonal properties of Chebyshev suits them better for such applications because unlike geometric moments their invariants can be calculated independently to arbitrary high orders without having to recalculate low order invariants [5-8]. These orthogonal properties also allow one to test up to what order to calculate these moments to get a good descriptor for a given database.

The reconstruction of Chebyshev moments can be computed as:

\[
f(x, y) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} T_{mn} t_m(x)t_n(y)
\]

The image reconstruction error provides a measure of the feature representation capability of the moment functions as shown in Figure 5. We use MSE as measure the difference between the reconstructed image and original image as in Figure 6. The symmetry property of Chebyshev polynomials can be made use to considerably reduce the time required for computing the associated moments.

Figure: 3. Example of frames that embedded a watermark of background subtraction of walking action of KTH.

The cumulative motion images (MEI) of six actions of KTH dataset starting from Frame 0.

Figure: 4. Cumulative motion images (MEI) of six actions of KTH dataset starting from Frame 0.
The accuracy of image reconstruction can be expectably better than the conventional continuous orthogonal moments such as Zernike moments.

[2.2] CLASSIFICATION METHOD

Multi-class SVM is used to classify the feature vector of the prototype video frames and the feature vectors stored in the KTH dataset.

[3] EXPERIMENTAL RESULTS

To recognize our representation, we divide the ‘KTH’ dataset for each action frames into two disjoint sets to make the problem more challenge. One set, contains 16 frames of each action [30], and is used for training, that embedding a watermark as human identity. The other set contains 9 frames and is used for testing. The frames are of dimension 160 x120, normalized to 64x64. A watermark identity image is first resized into small binary image with a size of 16x16 and it is embedded into different frames for each action video performed by distinctive person [8]. The procedure of embedding a watermark can be generalized to any video to identify the human that makes an action, so we do need any extra computational complexity for this problem. In general, we are training the action with identity at same time. To extract the watermarking by using inverse discrete wavelet transform (IDWT), this does not need any complexity computation. Chebyshev moment is used to extract features for action recognition. So we ended up with inputs of dimension as feature vector 78 of order 4. The extracted features from Chebyshev moments are used to train multi-class SVM [31]. We trained SVM for each separately as one vs all, which embedded a watermark in it, to learn the features of videos belonging to different classes and labeling of identity, and action labels are done in this step. In order to determine the best order of Chebyshev moment that is required for image representation without redundancy information, increasing order moments were used to reconstruct an object until the error between the original and reconstructed images was below a predefined threshold. For each class 100 frames having different styles of the same action performed by different person are taken for training. The only failed in identity recognition may be happened if the two persons or more are very similar in making the same action (same feature vectors), this sometimes appears in KTH dataset in indoor scenario because silhouette environment. Confusion between walking and jogging as well as between jogging and running can partly be explained by high similarities of these classes (running of some people may appear very similar to the jogging of the others). The confusion matrix of action recognition is 95%, and confusion matrix of identity recognition is 89.7% can be shown in Figures 7 and 8 respectively. The comparison of different methods can be shown in table 1.
Figure: 6. MSE comparison of images reconstructed using Chebyshev and Zernike moments

<table>
<thead>
<tr>
<th>Action</th>
<th>Walk</th>
<th>Run</th>
<th>Jog</th>
<th>H.w ave</th>
<th>H.clap</th>
<th>box</th>
</tr>
</thead>
<tbody>
<tr>
<td>walk</td>
<td>100.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>run</td>
<td>2.0</td>
<td>95.0</td>
<td>5.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>jog</td>
<td>3.0</td>
<td>2.0</td>
<td>95.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H.wave</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>94.0</td>
<td>0</td>
<td>6.0</td>
</tr>
<tr>
<td>H-clap</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.0</td>
<td>92.0</td>
<td>3.0</td>
</tr>
<tr>
<td>box</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.0</td>
<td>3.0</td>
<td>94.0</td>
</tr>
</tbody>
</table>

Figure: 7. Confusion Matrix of action accuracy used Chebyshev moment

<table>
<thead>
<tr>
<th>Identity</th>
<th>Jhon</th>
<th>Ali</th>
<th>Michel</th>
<th>Nar</th>
<th>Bob</th>
<th>Musa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jhon</td>
<td>90.0</td>
<td>5.0</td>
<td>5.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ali</td>
<td>0</td>
<td>88.4</td>
<td>7.2</td>
<td>4.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Michel</td>
<td>4.2</td>
<td>0</td>
<td>93.4</td>
<td>0</td>
<td>3.5</td>
<td>3.1</td>
</tr>
<tr>
<td>Nar</td>
<td>0</td>
<td>9.0</td>
<td>0</td>
<td>87.0</td>
<td>0</td>
<td>2.0</td>
</tr>
<tr>
<td>Bob</td>
<td>5.0</td>
<td>5.0</td>
<td>0</td>
<td>0</td>
<td>90.0</td>
<td>0</td>
</tr>
<tr>
<td>Musa</td>
<td>0</td>
<td>3.0</td>
<td>0</td>
<td>5.0</td>
<td>2.6</td>
<td>89.4</td>
</tr>
</tbody>
</table>

Figure: 8. Confusion Matrix of identity accuracy used watermark
Action Recognition And Identity Using Chebyshev Moments

<table>
<thead>
<tr>
<th>Method</th>
<th>Size</th>
<th>Computation time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>78</td>
<td>0.0233 s</td>
<td>95.0%</td>
</tr>
<tr>
<td>Zernike 2D [1]</td>
<td>54</td>
<td>0.0334 s</td>
<td>92.88%</td>
</tr>
<tr>
<td>Pyramid Zernike 3D [16]</td>
<td>84</td>
<td>0.0300 s</td>
<td>91.30%</td>
</tr>
<tr>
<td>Gradient + PCA[23]</td>
<td>100</td>
<td>0.0060 s</td>
<td>81.17%</td>
</tr>
<tr>
<td>3D SIFT [19]</td>
<td>640</td>
<td>0.8210 s</td>
<td>82.60%</td>
</tr>
<tr>
<td>3DGradn[21]</td>
<td>432</td>
<td>0.0400 s</td>
<td>90.38%</td>
</tr>
<tr>
<td>HOG-HOF [22]</td>
<td>162</td>
<td>0.0300 s</td>
<td>91.80%</td>
</tr>
<tr>
<td>HOG 3D [20]</td>
<td>380</td>
<td>0.0020 s</td>
<td>91.40%</td>
</tr>
<tr>
<td>SURF 3D [18]</td>
<td>384</td>
<td>0.0005 s</td>
<td>84.26%</td>
</tr>
<tr>
<td>ASIFT with identity[3]</td>
<td>200</td>
<td>----</td>
<td>81.2%</td>
</tr>
</tbody>
</table>

Table 1. Descriptor complexity comparison together with accuracy

Our results indicate that there is a low complexity computation to recognize actions and identity at the same time; we investigate to apply the proposed algorithm for real time video, where time is very critical factor.

[4] CONCLUSION

In this paper, we provide a new and simple solution to action recognition and identity at the same time. We have proposed a novel action recognition and identity framework by employing a watermark in the video, that robustly capture underlying human identity data, then using motion energy image to capture motion regions. The idea of implementing watermark embedding in the training phase is that poses as useful information about human identity. We use MEI as a feature representation of the human action, then discrete Chebyshev moments to extract features from MEI. The proposed method can automatically recognize actions and identity at the same time. The advantage of using Chebyshev moments than Zernike moments is discrete property which is suitable for image processing, so no quantization error of continuous integrals problems and robust to translation, rotation, scaling and noisy. The measure of feature vector descriptors of computed Chebyshev moments is by reconstruction error, which is a measure of a good representation of actions. Compared to the exist methods, it can meet high accuracy recognition by using only very small number of labeled training data. The future work is to prove a complex dataset for dynamic background such that Hollywood data sets and for complex actions.

[5] ACKNOWLEDGEMENT

The authors are grateful to the anonymous reviewers for their constructive comments which helpful to improve this paper.
REFERENCES


Action Recognition And Identity Using Chebyshev Moments


