Abstract—Human action recognition is a quite popular yet challenging problem in computer vision discipline, especially in automatical human motion understanding. This paper introduces a novel approach based on the second generation Curvelet transform to get the eigenvector for representing the human action in static images. As an exceptional multi-resolution feature extraction technique, the second Curvelet transform offers enhanced directional and edge representation that shows nice competitiveness. During feature descriptor extraction, the silhouettes and texture statistical information are extracted from the coefficients as the edge and the texture features. All the extracted features are aligned as the hybrid eigenvector of a frame. Experimental evaluation is performed on the benchmark Weizmann database and a comparison with the other counterparts is made. Results show that our method is rather competitive in quantitative index such as accuracy rate, which exhibits the descriptor developed from the second generation Curvelet to be a promising representation for such visual recognition tasks.

Keywords-component; Curvelet transform; edge features; texture features; human action recognition

I. INTRODUCTION

Despite the rapid pace of computer vision technologies, action recognition is still an appealing and challenging problem in research communities. Being an important part of human motion analysis, it leads to many applications such as advanced human-computer interaction, security surveillance, events analysis and virtual reality. However, how to recognize human actions is a tricky and prevalent problem even till the recent years. In this paper, we focus on the problem of human action recognition under uncontrolled monocular camera condition. Here we propose a method with a set of hybrid features of human motion that captures anisotropic lines and edges, so directional representations and as provides sparse coding solution is achieved in the meantime.

The current action recognition approaches range from constructing dynamic models to the spatio-temporal templates, most of which proved to be working well. The model-based methods fully leverage the prior knowledge of body structure to build models of the human and then extract the low-level features to match models, while in the meantime, it was highly constrained by motion rules. Existing models mainly include Hidden Markov models [1, 2] and Conditional random fields [3]. Also, J.C. Niebles and L. Fei-Fei presented a hierarchical model of shape and appearance for human action categorization [13]. Such modeling method relies on describing the details of the action dynamics and it works in a natural way. However, there is special necessity to deal with a large amount of training data to conduct an effective and precise model, which is the primary deficiency. Bobick and Davis again made the spatio-temporal templates popular[4], they introduced MEI and MHI to differencing the actions. What’s more, Biologically-inspired network [5, 6, 7] has been applied to action recognition tasks [19]. Recently, a novel action recognition approach with hierarchical invariant spatio-temporal feature-learning by independent subspace analysis is proposed [8].

As a fully action recognition system in general, it needs to tackle with three major subtasks: human detection, feature extraction and classification. The common seen classical descriptors for static image include HOG [9], SIFT [10], optical flow [14], SURF [11], and GLOH [12], etc. The extending descriptors for video include HOG3D, Cuboids, Harris3D, and so on. All these descriptors have respective strengths and weaknesses. Common technique that is used to extract features is linear filtering, but objects at different scales can arise from distinct physical processes [21]. Such approach further evolve into the use of scale-space filtering and multi-resolution wavelet transform.

In this paper, we introduce a novel multi-resolution feature extraction technique to recognize the multifold intricate actions that evolve from the second Curvelet transform coefficients as well as statistical silhouettes and texture information. The idea fixed on the Curvelet inherent attribution makes it a promising image representation. Upon the emergence of Curvelet, researchers have applied it in many image processing fields, such as denoising, seismic imaging, image fusion, texture or character classification, and face recognition [17]. Compared with other representation approaches, Curvelet offers enhanced directional and edge representation that shows compelling competitiveness.

Experiments are performed on the dataset of Weizmann. We found out more competitive performance compared with other state-of-art counterparts that worked frame-by-frame. The contribution of our work is summarized as follows: (1) A novel image representation approach to better depict image context is proposed, and shows promising results. (2)The operation of background subtraction can be omitted, which is computationally relaxed and time saving. (3)The utilization of the statistic features from the Curvelet coefficients reduces...
dimensionality, and also leads to a concise computational concept.

II. CURVELET TRANSFORM

One of the primary tasks in computer vision, especially in action recognition, is to extract effective features with selectivity from images. The features should be characterized by position, direction, scale, textures or other identifiable property parameters [21]. As a prospective multi-scale analysis method, Curvelet transform enables the characteristics of multi-scale and multi-directional. Thus, its special advance lies in representing the detailed curves and edges in the images, being more favorably for analyzing the curve-like context than conventional wavelet analysis. Also, the multi-scale character can help it detect image edges with different fineness and provide more context information in detailed description. As a kind of multi-resolution analysis, it also allows for zooming in and out of the edge structure, thus satisfies the needs for scale invariant in action recognition. Outline features extracted with Curvelet provide a good foundation for behavior analysis, where the second generation Curvelet theory also makes the theory easier to understand and implement.

In order to present the singularity of the curve efficiently, E.J. Candès provided an implementation of mono-scale ridgelets, and on which basis he and D.L.Donoho constructed a multi-scale ridgelet system named Curvelet [15]. Unlike Ridgelet, Curvelet has a variable length besides a variable width. In the finest scale, the width is the square of the length. That is \( \text{width} = \text{length} \times 2 \). Therefore, the anisotropic of Curvelet increases with the decrease of the scales. The core particularity of Curvelet is its capability of sparse representation by reducing the redundancy of scales. This particular step can be realized by multi-scale sub-band filter and orthogonal decomposition.

The difference between the first generation and the second generation of Curvelet lies in the theory and digital implementation. The formation idea of the first generation is to make sure the curve that approximate the line in each block to be small enough, and then analyze characteristics using local Ridgelet. The implementation procedure needs Sub-band Decomposition, Smooth Partitioning, Renormalization and Ridgelet Analysis. The entire notion is very complex and also brings about large data redundancy. By contrast, the second Curvelet has no need for complex Ridgelet, while it’s more succinct, faster and precise.

The second Curvelet transform [16] is defined directly via frequency partitioning. Defined in a polar coordinate system, the combination of angular and radial windows in frequency domain is called a ‘Curvelet’. Concentric circles break down the image into multiple scales and angular divisions are equivalent to disparate directions. As for Ridgelet, there are also digital Curvelet transforms. Candès etc constructed the tight frame for Curvelet and demonstrated that for those goal functions that have the singularity of the smooth curve, the reconstruction system of the digital Curvelet can provide a stable, efficient, and most importantly an almost perfect representation.

For digital Curvelet transform, there are two separate algorithms, Unequally-Spaced Fast Fourier Transforms (FDCT_USFFT) and Wrapping of specially selected Fourier samples (FDCT_WARP). The former achieves the coefficients by irregularly sampling, while the latter gets the coefficients by a series of translations and a wraparound skill. Although the same result can be achieved, the Wrapping algorithm is intuitive and timesaving [18]. We adopt the Wrapping algorithm in our experiments. Anyway, the Curvelet transform can not only represent the edge that has discontinuity but also perform well.

It is a multi-scale transform with tight frame where elements are indexed by scale parameters, orientation parameters \( \theta \) and location parameters \( b \), such as \( \mu = (a, \theta, b) \). In which, \( a (0 < a < 1) \) is the scale parameter. More details can be found in [21].

III. FEATURE EXTRACTION

The new method that we propose for human action recognition is based on the second Curvelet transform and makes use of the concept of GLCM. The fusion of both information working for the feature extraction can describe a image strictly. The detailed procedures of the proposed algorithm are as follows:

A. Curvelet Coefficients

There is no need for pre-processing for still action images. Initially, we briefly set the parameter. By means of the Curvelet transform tools [20], the coefficients of all sub-bands can be obtained in matrices. A visual exhibition is given in Fig.1.

The visual display of the Curvelet coefficients.

![Visual display of Curvelet coefficients](image)

For digital Curvelet transform, there are two separate algorithms, Unequally-Spaced Fast Fourier Transforms (FDCT_USFFT) and Wrapping of specially selected Fourier samples (FDCT_WARP). The former achieves the coefficients by irregularly sampling, while the latter gets the coefficients by a series of translations and a wraparound skill. Although the same result can be achieved, the Wrapping algorithm is intuitive and timesaving [18]. We adopt the Wrapping algorithm in our experiments. Anyway, the Curvelet transform can not only represent the edge that has discontinuity but also perform well.

It is a multi-scale transform with tight frame where elements are indexed by scale parameters, orientation parameters \( \theta \) and location parameters \( b \), such as \( \mu = (a, \theta, b) \). In which, \( a (0 < a < 1) \) is the scale parameter. More details can be found in [21].

III. FEATURE EXTRACTION

The new method that we propose for human action recognition is based on the second Curvelet transform and makes use of the concept of GLCM. The fusion of both information working for the feature extraction can describe a image strictly. The detailed procedures of the proposed algorithm are as follows:

A. Curvelet Coefficients

There is no need for pre-processing for still action images. Initially, we briefly set the parameter. By means of the Curvelet transform tools [20], the coefficients of all sub-bands can be obtained in matrices. A visual exhibition is given in Fig.1.

The visual display of the Curvelet coefficients.

![Visual display of Curvelet coefficients](image)

For digital Curvelet transform, there are two separate algorithms, Unequally-Spaced Fast Fourier Transforms (FDCT_USFFT) and Wrapping of specially selected Fourier samples (FDCT_WARP). The former achieves the coefficients by irregularly sampling, while the latter gets the coefficients by a series of translations and a wraparound skill. Although the same result can be achieved, the Wrapping algorithm is intuitive and timesaving [18]. We adopt the Wrapping algorithm in our experiments. Anyway, the Curvelet transform can not only represent the edge that has discontinuity but also perform well.

It is a multi-scale transform with tight frame where elements are indexed by scale parameters, orientation parameters \( \theta \) and location parameters \( b \), such as \( \mu = (a, \theta, b) \). In which, \( a (0 < a < 1) \) is the scale parameter. More details can be found in [21].
meaning underlying features, statistical measurements are used as discriminative features.

Inspired by this, we employ the extraction approach as follow:

- Every scale is divided into a set of blocks in a non-overlapping manner. Because the coarsest scale in image level possesses the most information, so the first sub-band is partitioned by 4*4 sized grid, with the remaining sub-bands partitioned by 8*8. See Fig.2.

- Beyond each block, calculate the statistical elements which include its energy (ENG) and contrast (CON), together they form the edge feature vector. The energy and contrast as local descriptor describe the strength of the margin information and the space distributive condition of the margin respectively. Their formulations are shown below:

  \[
  ENG = \sum_{i,j} c_{i,j}^2 .
  \]

  \[
  CON = \sum_{i,j} (i - j)^2 c_{i,j} .
  \]

Where, \( c_{i,j} \) is the coefficient in each block, with the location \((i, j)\).

- Normalization.
  For ENG,

  \[
  V_{eng\_norm} = \frac{V_{eng} - \min V_{eng}}{\max V_{eng} - \min V_{eng}} .
  \]

  For CON,

  \[
  V_{con\_norm} = \frac{V_{con}}{\max \{abs V_{con}\}} .
  \]

The edge feature vector in block is \( V_{edge} \).

- Concatenating the above features into the final edge feature vector.

  \[
  V_{edge} = \{ENG, CON\} .
  \]

The whole edge feature extraction procedure is shown in Fig.2. The top line picture is a gray image in the Weizmann dataset. The middle level of picture represents the five sub-band coefficients. The remaining details indicate how the feature is computed.

C. Texture feature extraction

Initially Haralick etc. put Gray-level Co-occurrence Matrix (GLCM) forward to depict the texture statistic feature in 1973. On account of the ability to represent the object surface, it results in extensive applications. Here, we take advantage of the GLCM concept to compute the co-occurrence matrix of the coefficients of sub-bands in transform domain as the texture features [22].

Because of the anisotropy attribute and the aim at utilizing the multi-directions and angle characteristic, we employ different feature extraction strategies in dissimilar sub-bands.

For the first sub-band, we compute the co-occurrence matrix of the coefficients by dividing the whole coefficients matrix into 4*4 blocks. The assigned blocks overlap each other by 1/2. The co-occurrence matrix is generated in four orientations \((0, 45, 90, \text{and } 135)\) of each block and its discrete series is 16.

For the coefficients co-occurrence matrix, six statistic measures (including ASM EMP CON COR SOA and SOV) are computed, which are composed of the texture feature and present the texture information adequately. Hereinto, CON and COR depict the degree of texture and sharpness; SOV describes the size of the texture’s period; ASM and EMP express the degree of thickness and uniformity; SOA gives the levels of the total tonal. They are calculated by formulas below.

\[
ASM = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p(i,j)^2 .
\]

\[
CON = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (\sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p(i,j))^2 .
\]

\[
COR = \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} (p(i,j) - \mu_x \mu_y) / \sigma_x \sigma_y .
\]

Where \(|i - j| = n\).
\[ CON = \sum_{k=1}^{G-1} \left( \sum_{i=0}^{G-1} \sum_{j=0}^{G-1} p(i,j) \right), \] (10)

Where,

\[ \mu_i = \sum_{k=1}^{G-1} \sum_{j=0}^{G-1} p(i,j), \quad \mu_j = \sum_{k=1}^{G-1} \sum_{j=0}^{G-1} p(i,j). \]

\[ \sigma^2 = \sum_{i=0}^{G-1} (j-\mu_j)^2 \sum_{j=0}^{G-1} p(i,j), \quad \sigma^2 = \sum_{i=0}^{G-1} (j-\mu_j) \sum_{j=0}^{G-1} p(i,j). \]

\[ SOA = \sum_{i=1}^{G} k \hat{p}_{xy}(k). \] (11)

Where, \( \hat{p}_{xy}(k) = \sum_{i=1}^{G} \sum_{j=1}^{G} \hat{p}(i,j), \quad k = 2, 3, ..., 2G \)

\[ SOV = \sum_{i=1}^{G} (k - SOA)^2 \hat{p}_{xy}(k). \] (12)

\[ ENP = -\sum_{i=1}^{G} \sum_{j=0}^{G-1} p(i,j) \log \hat{p}(i,j). \] (13)

In the above equations, \( \hat{p}(i,j) \) is the element in the partitioned block with location \((i, j)\). G is the quantization series of the co-occurrence matrix. Being the same as the construction as the edge feature, the texture feature vector is

\[ V_{\text{texture}} = \{ ASM, EMP, CON, COR, SOA, SOV \} \] (14)

The normalization is conducted through the following expression.

\[ V_{\text{texture}} = \frac{V_{\text{texture}} - \min V_{\text{texture}}}{\max V_{\text{texture}} - \min V_{\text{texture}}} . \] (15)

Finally, concatenate the normalized texture feature vectors of blocks into a generally fixed texture feature vector. Just as

\[ V_{\text{texture}} = \bigcup_{q=1}^{k} V_{\text{texture}} . \] (16)

Where, \( k \) is the number of the dissection units.

The extraction of the texture feature can be seen in Fig.3, where the left part expresses the operation of dividing the coefficients matrix into blocks in S1 and the right one is the visualized distribution of the final texture feature vector. The ultimate eigenvector that depicts the action body is the hybrid feature by incorporating the edge feature and texture feature.

IV. EXPERIMENTAL RESULTS

To validate the efficiency of our approach, we use the Weizmann dataset here introduced by Blank et al., which is one of the primary benchmark datasets in human action recognition field.

A. Database

Weizmann dataset [24] contains ten actions: walk, run, jump, gallop sideways, bend, one-hand wave, two-hand wave, jump in place, jumping jack and skip. Except the actions ‘walk’ ‘run’ ‘skip’ that are acted by ten persons, the rest of the actions are implemented by nine persons. The total number of videos is 93.

Fig.4 is the example sequences of the ten actions.

Fig.5. The action ‘jack’ did by nine persons.

In Fig.5, the action ‘jack’ is given independently by nine different persons.

Finally, concatenate the normalized texture feature vectors of blocks into a generally fixed texture feature vector. Just as

\[ V_{\text{texture}} = \bigcup_{q=1}^{k} V_{\text{texture}} . \] (16)

Where, \( k \) is the number of the dissection units.

The extraction of the texture feature can be seen in Fig.3, where the left part expresses the operation of dividing the coefficients matrix into blocks in S1 and the right one is the visualized distribution of the final texture feature vector. The ultimate eigenvector that depicts the action body is the hybrid feature by incorporating the edge feature and texture feature.
B. The Experimental Setup

The whole experimental procedure is displayed in Fig.6. Based on the Weizmann dataset, we construct the training set and the test set by leave-one-out method, then pick up the frames of each action video to constitute the traindata set andtestdata set. There are totally 5687 frames in dataset, 5161 frames for training and 526 for testing, with all size of 180*144. For each of the frames in dataset, the hybrid feature is extracted by the proposed method. Accordingly the train data and test data set come into being. The extracted feature is 1568 dimension, including the 1472 dimensional edge feature and the 96 dimensional texture feature.

The ten actions ‘bend’ ‘jack’ ‘jump’ ‘pjump’ ‘side’ ‘wave1’ ‘wave2’ ‘run’ ‘walk’ ‘skip’ are labeled ‘1’ ‘2’ ‘3’ ‘4’ ‘5’ ‘6’ ‘7’ ‘8’ ‘9’ ‘10’ respectively. By Considering 10 actions involved, we construct a multi-classifier consisted of 45 binary classifiers by utilizing the linear SVM in libSVM.

C. Results Analysis

For the parameter setting, the accuracy rate of cross validation process is 94.59%. The confusion matrix of ten actions is displays in Fig.7. The total classification accuracy is 83.0798%. As we can see, most of the results are satisfying, where even the accuracy of ‘one-hand wave’ reaches 100%. The visual similarities among the action ‘walk’ ‘run’ and ‘skip’ inherently results in confusions or ambiguity.

From Fig. 8, which shows some parts of the ROC curves corresponding to different actions, it reveals the higher recognition rate and lower false positive rate, especially for the action “jack”.

Nevertheless, if we choose the nine actions except ‘skip’ as described in [13], the comparative result can be shown below and the accuracy is 87.55%, being obviously superior to 72.8% reported in [13].

Figure 7. The confusion matrix of ten actions.

Figure 8. Parts of the ROC curves for corresponding actions.

Figure 9. The contrast of the confusion matrixes between ours (top) and [13] (bottom).
The vertical columns are the predicted labels, and the horizontal rows show the ground truth labels. From the confusion matrices, it is obvious that most of the actions are precisely recognized. From here, we can come to the conclusion that the hybrid features extracted by our approach can describe the pose excellently and they reveal strong selectivity power to express the silhouette or margin. However, in Fig. 7, ‘run’, ‘walk’ and ‘jump’ give unsatisfactory result, mostly because of their subtle visual differences to ‘skip’. This is also on account of the special-temporal information we have not yet included. Hence, we are planning to conduct broad experiments on the other databases for more contrast tests, and also attending to utilize the temporal features to describe especially the easily confused actions.

V. CONCLUSION

To our best knowledge, this paper addresses for the first time the utilization of the hybrid feature developed from the second Curvelet transformation as the descriptor for action recognition. Among the features from sparse coding, the edge features are extracted through incorporating the statistical information that come from equally-sized block. And the co-occurrence features are extracted from the directional co-occurrence matrixes of the blocks. Such sparse directional representation motivated by Curvelet ensures strictly localization of edges and curves with multi-scale, rotation invariance, as well as directional selectivity for complex subjects such as human-beings ourselves.

Given a multitude of merits based on geometry of the second Curvelet theory, the results show that the proposed method is promising and computationally applicable, especially when we reinvent the Curvelet transform for the multi-resolution analysis method. Also, the anisotropy and directionality characteristics greatly help representing human actions robustly and sparsely. In our future work, we will endeavor to take advantage of the inter-frame information to construct features based on temporal relations, so that it’s hoped to promote the efficiency of feature description as well as the accuracy of classification.

REFERENCES


