ABSTRACT

This paper presents a simple and efficient method for action recognition based on the learning of an explicit representation for an intrinsic dynamic shape manifold of human action. The proposed model allows for a fast spatio-temporal dimensionality reduction technique for embedding a sequence of 2D images into a low-dimensional space. The intrinsic dynamic shape manifold of the actions, which is defined as a manifold of the action skeleton in a shape space, is characterized by the correlation of the silhouette sequences. Moreover, we demonstrate the efficiency of our method in partial occlusion, deformation of shapes, significant changes in scale and viewpoint, irregularities in the performance of an action, and low-quality video.

Introduction

- Human activity recognition:
  - Local representations
  - Global representations:
    - Spatio-temporal template matching
    - Manifold learning based

- Our solution: hybrid-framework which combines both manifold learning and spatio-temporal template matching technique

Spatio-Temporal Action Volume (STV)

Build a centered motion field of a moving body by aligning the 2D center of mass corresponding to each pixel-vector contains the intensity values that the pixel takes over the torso of the person over time.

- To treat both periodic and non-periodic actions and to compensate for different lengths of the sequences:
  - Use a temporal sliding window of 10 frames with an overlap of 5 frames along the temporal axis.

- Divide STV into \( H \times W \) vectors
  - \((H, W)\) are, respectively, height and width of the image
  - Each pixel-vector has a dimension \( T \) (number of frames in the STV)
  - Each pixel-vector contains the intensity values that the pixel takes over \( T \) consecutive frames of a video sequence.

- FastMap mapping from TD to 1D along the temporal axis
  - Each pixel-vector in the initial space corresponds to a point in the reduced space
  - These points will represent the pixel intensity values of the STAS image

Spatio-Temporal Action Shapes (STAS)

- Mean correlation rates for the FastMap reduction technique:
  - estimate the correlation \( \rho \) of the Euclidean distance between each pairwise vectors-pixels in the high dimensional space (let \( X \) be this vector) following equation:

\[
\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}
\]

where \( X, Y \), \( X \) and \( Y \) respectively represent the transpose, cardinality, mean and standard deviation of \( X \).

<table>
<thead>
<tr>
<th>Actions</th>
<th>Kick</th>
<th>Jack</th>
<th>Jump</th>
<th>Pjump</th>
<th>Run</th>
<th>Side</th>
<th>Skip</th>
<th>Walk</th>
<th>Wave1</th>
<th>Wave2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>0.19</td>
<td>0.88</td>
<td>0.95</td>
<td>0.94</td>
<td>0.89</td>
<td>0.90</td>
<td>0.90</td>
<td>0.97</td>
<td>0.92</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Conclusion

- Our method:
  - simple, inexpensive and fast allowing simple action recognition.
  - does not require video alignment, and is applicable in many situations.
  - is robust to partial occlusion, deformation of shapes, significant changes in scale and viewpoint, irregularities in the performance of an action, and low-quality video.

References


Sample images representing the different action STAS

Experimental Results

- Use a Nearest Neighbors Classifier with leave-one-out procedure to label the test actions
- Recognition results on the Weizmann dataset:

<table>
<thead>
<tr>
<th>Method</th>
<th>Correlation</th>
<th>SVM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>90.7%</td>
<td>99.1%</td>
<td>99.1%</td>
</tr>
<tr>
<td>SVM</td>
<td>98.3%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>99.0%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>99.0%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>99.0%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Correlation</th>
<th>SVM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>90.7%</td>
<td>99.1%</td>
<td>99.1%</td>
</tr>
<tr>
<td>SVM</td>
<td>98.3%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>99.0%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>99.0%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>99.0%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Correlation</th>
<th>SVM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>90.7%</td>
<td>99.1%</td>
<td>99.1%</td>
</tr>
<tr>
<td>SVM</td>
<td>98.3%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>99.0%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>99.0%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
<tr>
<td>SVM</td>
<td>99.0%</td>
<td>99.8%</td>
<td>99.8%</td>
</tr>
</tbody>
</table>

Recognition results on the KTH dataset:

- Our method Failure 100%

References