Human Action Recognition using Segmented Skeletal Features

Sang Min Yoon
GRIS, TU-Darmstadt
sangmin.yoon@gris.informatik.tu-darmstadt.de

Arjan Kuijper
Fraunhofer IGD & TU-Darmstadt
arjan.kuijper@igd.fraunhofer.de

Abstract

We present a novel human action recognition system based on segmented skeletal features which are separated into several human body parts such as face, torso and limbs. Our proposed human action recognition system consists of two steps: (i) automatic skeletal feature extraction and splitting by measuring the similarity in the space of diffusion tensor fields, and (ii) multiple kernel Support Vector Machine based human action recognition. Experimental results on a set of test database show that our proposed method is very efficient and effective to recognize human actions using few parameters, independent of dimensions, shadows, and viewpoints.

1 Introduction

Human action analysis and recognition systems, defined to understand the basic human actions such as jogging, walking and boxing from images, has a long history in the area of computer vision and pattern recognition. Interest in this topic is encouraged by a multitude of applications, such as automated surveillance systems, smart home applications, video indexing and browsing, virtual reality, human-computer interaction and analysis of sports events. Human action recognition from a single 2D image is heavily studied by numerous researchers [1-6], but still a challenging issue due to partial occlusion, clutter, dependence of viewpoint and pose ambiguity within a 2D image.

An important issue of human action recognition systems, which can be interpreted as one kind of object recognition and retrieval systems, is how to define the appropriate similarity measure from reliable features of deformable objects that have a high-degree of freedom, and to automatically assess the similarity between any pair of human motions based on a suitable notion of similarity.

We propose to solve these problems of human action recognition by focusing first on the adequate automatic skeletal feature extraction and separating human body model into several human body parts like head, torso and limbs, because skeleton based object recognition systems generally perform better than shape based object recognition approaches [7]. We extract and split the human skeleton using Normalized Gradient Vector Flow in the space of diffusion tensor fields, using the eigenvalues and eigenvectors of the segmented skeletal features (see Sec. 2.1). Second, we adapt a multiple-kernel Support Vector Machine approach (see Sec. 2.2) for the actual human action recognition. Figure 1 shows the total flowchart of our proposed method.

1.1 Human action recognition systems

Human action recognition systems can be largely separated into four categories. First, structural methods use parameterized models describing geometric configurations and relative motions of parts in the motion patterns. Second, appearance-based methods using template features need a lower degree of freedom than those of structural approach. The statistical approach was proposed to overcome the difficulty of finding cor-
responding features between models and structures in test images of structural and appearance based methods. Lastly, the event-based motion interpretation method are popularly used for human action recognition.

2 Our Approach

Most for human action recognition systems are based on computing local space-time gradients or other intensity based features. It might be unreliable in the cases of low quality video, motion discontinuities and motion aliasing. To overcome these problems, we will explain our new approach for skeletal feature extraction and splitting, similarity measure for action classification and recognition.

2.1 Automatic skeletal feature extraction and splitting

A precise definition of the skeleton or medial axis (MA) in the continuum was started by Blum [8]. It is a compact one dimensional representation of complex and deformable objects. It also describes an object’s geometry and topology using little data.

Previous skeleton extraction algorithms are often not optimally performing because of their high computational complexity, noise sensitivity, centeredness inside the underlying complex shape, partial occlusion, or artifacts in a singular region of the given shape. The skeleton extraction using Normalized Gradient Vector Flow (NGVF) fields in the space of diffusion tensor fields overcomes the disadvantages of previous works [9]. Each pixel in the image is represented by each tensorial elements as follows:

\[
T(\bar{x}) = \begin{pmatrix}
T_{11}(x,y) & T_{12}(x,y) \\
T_{21}(x,y) & T_{22}(x,y)
\end{pmatrix}
= \lambda_i(\bar{x})\bar{e}_i(\bar{x}),
\]

(1)

where \(i = 1, 2\). \(\lambda_i(\bar{x})\) are the eigenvalues of \(T(\bar{x})\) and \(\bar{e}_i(\bar{x})\) define the unit eigenvectors. The skeleton is extracted by connecting and thinning the degenerate points, obtained when the two eigenvalues of \(T(\bar{x})\) are equal to each other.

The extracted skeleton is split by measuring the similarity between neighboring pixels. The segmentation of the extracted skeleton is computed by measuring the similarity between neighboring pixels.

For each pixel \(I_i\) which is recognized as a skeleton point, we measure the dissimilarity between neighboring skeleton elements using tensorial dissimilarity function. Given two tensors \(T_i\) and \(T_j\), there are some dissimilarity measures that might be used to compare them. The tensor can be represented by an ellipsoid, where the lengths of medial axis are proportional to the square roots of the tensor eigenvalues \(\lambda_1\) and \(\lambda_2\) (\(\lambda_1 > \lambda_2\)) and their direction correspond to the respective normalized eigenvectors. With these properties, we can measure the dissimilarity between neighboring elements. The simplest one is the tensor dot product:

\[
d_1(T_i, T_j) = \sum_{i}^{2} \sum_{j}^{2} \lambda_i^1 \lambda_j^2 (e_i^1 \cdot e_j^2)^2.
\]

(2)

A second dissimilarity measure that uses the full tensor information is the Frobenius norm:

\[
d_2(T_i, T_j) = \sqrt{\text{Trace}((T_i - T_j)^2)}.
\]

(3)

The dissimilarity measure between two elements is the multiplication of \(d_1\) and \(d_2\). The skeleton splitting methodology is determined by comparing the similarity measure between neighbor points when the direction of NGVF changes and the scale of the main and the sub eigenvalue is over a pre-set threshold. Figure 2 shows the derived skeleton and its ellipsoidal representation. The segmented skeleton which is colored by different colors is represented by averaging the eigenvalues and eigenvectors. The scale and rotation of each ellipsoidal model are determined by its averaged eigenvectors and eigenvalues. The input for classifying the human action is composed of the matrix of segmented skeletal features like \(x_i = \frac{1}{N_i} \sum_{n=1}^{N_i} e_n \cdot \lambda_n\) where \(N_i\) is the number of segmented skeleton pixels.
2.2 Multiple kernel SVM based human action recognition

In this section, we describe how the Support Vector Machine (SVM) is used for efficient classification of highly variant human motions. SVM supports the classification tasks and handles the multiple continuous and categorical variables. The performance of different classifiers applied in object detection and recognition systems have been evaluated and compared in the area of pattern recognition and data learning. Bazzani [10] concluded that the Support Vector Machine (SVM) performs better than the Multi-Layer Perceptron (MLP) for a small number of training data. Having evaluated the SVM, Kernel Fisher Discriminant (KFD), Relevance Vector Machine (RVM), Feedforward Neural Network (FNN), and committee machines, Wei, et al. [11] concluded that the Kernel-based classifications like SVM yielded the best performance.

In \( \{x_i, y_i\}_{i=1}^{l}, x \in \mathbb{R}^m \) where \( l \) the number of training features, each \( x \) is then mapped to a \( \Phi(x) \) and \( y_i \) is separated into human actions like boxing, jogging, and walking. The training features which are used in our approach are the averaged eigenvalues and eigenvectors.

The non-linear SVM maps the training samples from the input space into a higher-dimensional feature space via a mapping function \( \Phi \) and construct a hyperplane defined as \( w^T \Phi(x) + b = 0 \) to separate examples from the classes. \( \{x_i, y_i\}_{i=1}^{l} \) in the kernel-induced feature space is related to the kernel function \( K \) which intuitively computes the similarity between examples in SVM. The standard SVM [12] tries to find a hyperline that has large margin and small training error. Instead of having a single kernel (SK) \( K \), suppose that we have a set of \( M \) base kernels \( K_1, K_2, ..., K_M \) with corresponding kernel-induced feature maps \( \Phi_1, ..., \Phi_M \). The MK-SVM [11] is extended from the SK-SVM as follows:

\[
\begin{align*}
\min_{w, b, \xi} & \frac{1}{2} \sum_{k=1}^{M} ||w_k||^2 + C \sum_{i=1}^{l} \xi_i \\
\text{s.t.} & \quad y_i (\sum_{k=1}^{M} w_k^T \Phi_k(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, ..., l, \quad \text{where} \quad w = \{w_1, w_2, ..., w_M\} \text{ which is the weight for component } \Phi_k, \text{ and } \xi \text{ is the non-negative slack variables. The regularization parameter } C \text{ determines the trade-off between the maximization margin } \frac{1}{||w||^2} \text{ and the minimum experience risk.}
\end{align*}
\]

The eigen-features \( x \) extracted from segmented human body parts (see previous section) are used for the learning based human action recognition system. They represent the characteristics of the segmented subregion of human body and are therefore used for the classification of the human actions.

3 Experiments & Discussion

For human action recognition, we used public image data such as the HumanEva database\(^1\) and the KTH human action dataset\(^2\) to compare our method to other proposed human action recognition approaches. The HumanEva dataset is separated into boxing, walking, and jogging, while the KTH human action database is more specifically separated into 6 human actions like hand-clapping, hand-waving, jogging, running, walking, and boxing. In the Multiple Kernel Support Vector Machine methodology, various kernels like Radial Basis Function (RBF), quadric, and linear kernels are used for a robust action recognition system.

Table 1 shows the human action recognition results of the HumanEva database. Table 1(a) shows the our proposed methodology from the 7 different viewpoints and Table 1(b) is the human action recognition using Conditional Random Fields for 6 models [6].

Unfortunately, it is difficult to directly compare the our

\(^1\)http://vision.cs.brown.edu/humaneva/
\(^2\)http://www.nada.kth.se/cvap/actions/
Table 2. 2D human action recognition ratio of the KTH Dataset using different classification methods

<table>
<thead>
<tr>
<th>Action</th>
<th>Run</th>
<th>Walk</th>
<th>Jog</th>
<th>Box</th>
<th>Hand-clap</th>
<th>Hand-wave</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang</td>
<td>0</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Roth et al. [2]</td>
<td>0</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Danafar et al. [3]</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Niebles et al. [4]</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Yeo et al. [5]</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(a) our approach (b) Wang et al. [1] (c) Roth et al.[2] (d) Danafar et al. [3]

The KTH human action data is less dependent on camera viewpoint than the HumanEva dataset, but its image resolution is less than the HumanEva dataset and has noise in background. The experimental results presented in Table 2 show that our proposed method (Table 2(a)) is again more balanced in recognizing the various human motions than other methods (Table 2(b-f)) which are based on local features, spatio-temporal features or optical flow [1-5]. This is because our method is very robust against noise, clutter, pose ambiguity and illumination. We also effectively classify the features by measuring the dissimilarity between actions using tensorial features.

Generally, there are many reasons for not successfully recognizing human actions, like shadows, viewpoints, and the ambiguity of human actions. In our approach, we were only not able to extract the meaningful features when the arms were located in the torso.

4 Summary & Outlook

In this paper we presented a novel human action recognition technique whose properties come from the segmented human body parts’ eigenvalues and eigenvectors, derived using diffusion tensor fields. These properties are used in a multiple kernel SVM approach yielding an efficient and effective human action recognition system. Experiments on publicly available data sets show that our method results in recognition rates that are better and more stable than those from other approaches.

Future work will focus on efforts to extract the robust features from a target object to effectively understand human behaviors, and extensions to 3D models.

References