Arm-hand Action Recognition Based on 3D Skeleton Joints

Ling RUI¹, Shi-wei MA¹, a, *, Jia-rui WEN¹ and Li-na LIU¹, 2

¹School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China
²School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China

a masw@shu.edu.cn
* Corresponding author

Keywords: Arm-hand action recognition, K-means, Bag-of-words, SVM, Kinect.

Abstract: In this paper, a simple and reliable method for arm-hand action recognition is proposed based on the representation of 3D skeleton joints. It mainly consists of two parts, feature extraction and classifier design. First, through the whole 3D skeleton joints collected by Kinect sensor, a hand joint, which is selected to represent the arm-hand action, is fed to a k-means cluster to decrease the number of features. Meanwhile, a data preprocessing of rotation is employed to deal with the multi-view problem. Then, the bag of words model is applied to form codebooks for each kind of arm-hand action. Finally, arm-hand action recognition is implemented by training SVMs using different kernel functions and parameters. The method was tested and compared with some state-of-the-art approaches on a self-built dataset for action recognition. The results manifested the effectiveness of the proposed method.

Introduction

Human action recognition including whole body motion, leg or hand motion has attracted a lot of attentions by many researchers due to the rapid development of computer vision and the demand of human computer interaction (HCI). Among all the human body structure, arm-hand may be the most flexible part whose motion exhibits a variety of key human action meanings. Hence, the image pattern recognition technology of arm-hand action becomes an active research topic. Nevertheless, because of the individual differences and self-occultation of arm-hand action performance, the research has become a challenging problem.

Early researches of arm-hand action usually resort to wearable devices such as data gloves which can achieve high accuracy [1]. But using the devices has equipment limitation and it is not interactive friendly. Instead, utilizing computer vision to recognize arm-hand action is a more natural way because of the low cost and the undemanding of external devices. Grif et al. proposed a new feature from color image recorded by video cameras to detect hand for controlling the mouse curse [2]. Although arm-hand action, hand gesture and human action recognition are three different sides of researches, they have similarities in methodology. Basically, there are two main steps during these recognition problems which are feature extraction and recognition. Traditional literatures about feature extraction mainly focus on features recorded from 2D cameras, such as silhouettes, edges, motion energy image(MEI)[3], motion history image(MHI) which are global features and Harris corner, interest point feature which are local features. All the features mentioned above are sensitive to noise, occlusions and variations in viewpoint.

Human action exists in 3D space during our real life, so 3D data can provide us much more information than the traditional 2D ones, which can be obtained from the cost-effective depth sensor recent years, such as Kinect. Li et al. found bag of spatiotemporal features called video-words by quantizing the extracted 3D interest points (cuboids) from the videos [4]. Liu and Shah adopted the expandable graphical model to explicitly model the temporal dynamics of the actions and used a bag of 3D points (BOPs) extracted from the depth map to model the postures [5]. In recognition aspect, there are some classical methods. Maierdan et al. trained a left-to-right Hidden Markov Model (HMM) for each kind of action to recognize different actions [6]. Gori et al.
used a linear Support Vector Machines (SVMs) to model a classifier and proposed an on-line algorithm to cope with the real-time recognition of primitive sequences [7].

Arm-hand action may refers to hand posture or hand gesture [8-9] which focus on palm part, while in our study, more attention is paid to the movements of arm including shoulder, elbow and hand. In this paper, a simple and reliable method by the combination of bag of words model and SVM for arm-hand action recognition based on the representation of 3D skeleton joints is proposed. The rest of the paper is organized as follows. Section 2 gives the method of data preprocessing to make the recognition system view invariant. Section 3 describes an overview of the arm-hand action recognition system and two main parts of the system. Experimental results and conclusions are showed in section 4 and section 5 respectively.

Data Preprocessing

Our proposed method base on the skeleton joints acquired from Kinect sensor which is shown in Fig.1. The data from Kinect sensor contain three streams which represent the 3D point (x, y, z) where X, Y and Z denotes as the horizontal axis, the vertical axis and the distance between a joint and the sensor, respectively. So with the help of the code provided by Microsoft can provide us 20 3D coordinate points for each action frame. In this paper, we mainly focus on one arm-hand action. Usually, the arm-hand actions to be recognized are the front ones. For these skeleton joints of actions, normalization is necessary to make them coordinate translation invariant. Since we study the motion of right hand and the right shoulder joint almost invariant, we can choose right shoulder joint as our reference joint. These three joints' X, Y and Z values are centered in the origin with reference to the reference joint position. Thus, no matter how far a person is from the Kinect sensor, it will be invariant with respect to distances. This can be realized through the following Eq.1- Eq.3.

$$X' = X - X_{ref}$$  \hspace{1cm} (1)

$$Y' = Y - Y_{ref}$$  \hspace{1cm} (2)

$$Z' = Z - Z_{ref}$$  \hspace{1cm} (3)

While not all the arm-hand actions are the front ones shown in Fig.2, there are some 45 degree angle action shown in Fig.3 including left 45 degree and right 45 degree. For these actions, choosing correct frame of reference can make the data view invariant and rotation transformation is

![Figure 1. 20 Human Joints achieved from Kinect view.](image1.png)

![Figure 2. Draw circle action in front view.](image2.png)

![Figure 3. Draw circle action in 45 degree view.](image3.png)
applied here. We redefine the line of right shoulder and left shoulder joint in X axis as new X axis of reference and calculate the angle between the original X axis and new axis called $\theta$. The $\theta$ is calculated by the arc-tangent ranging from $-90^\circ$ to $90^\circ$. Then all the skeleton joint are rotated along Y axis by $(-\theta)$ degree with equation (4) which translates the 45 degree view to the in front view.

$$
(x', y', z') = (x, y, z) \begin{bmatrix}
\cos(-\theta) & 0 & -\sin(-\theta) & 0 \\
0 & 1 & 0 & 1 \\
\sin(-\theta) & 0 & \cos(-\theta) & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
$$

(4)

After the rotation, normalization will also be used to make all skeleton joint not only view invariant but also translation invariant. As a consequence of this, raw data are transformed into new data which are prepared to feed to k-means cluster.

Proposed Method

In this paper, k-means cluster and bag of words model are used to extract feature for arm-hand actions and SVMs are then utilized to recognize different kinds of actions. Fig. 4 shows the schematic illustration of our proposed method.

![Figure 4. Overview of the proposed method.](image)

K-Means Cluster

According to the data reprocessing method mentioned above, we can get normalized skeleton joints. For the five actions recognition: high arm wave, draw circle, hand clap, forward punch, high throw, we have already fixed the right shoulder joint, and the trajectory of the right hand joint can separate each action category obviously. So the right hand joint is selected to be the key joint. In order to reduce the feature dimension, k-means algorithm is employed. With the given parameters: number of clusters $K$, cluster initialization, the following objective function is used:

$$
J(C) = \sum_{k=1}^{K} J(c_k) = \sum_{k=1}^{K} \sum_{i \in c_k} \| \omega_i - \mu_k \|^2 = \sum_{k=1}^{K} \sum_{i=1}^{n} d_{ik} \| \omega_i - \mu_k \|^2, \quad d_{ik} = \begin{cases} 1, & \text{if } \omega_i \in c_k \\
0, & \text{if } \omega_i \notin c_k \end{cases}
$$

(5)

$$
\| \omega_i - \mu_k \|^2 = (x_i - x_k)^2 + (y_i - y_k)^2 + (z_i - z_k)^2
$$

(6)

Here, $\omega_i$ is one of normalized joint data $(x_i, y_i, z_i)$, $\mu_k$ is the center of $k^{th}$ cluster $(x_k, y_k, z_k)$, $\| x_i - \mu_k \|^2$ is the Euclidean distance between $\omega_i$ and the nearest cluster center. Euclidean distance is chosen in the k-means cluster to translate the adjoining spatial joints into same cluster label in order to reduce the feature number. When using k-means algorithm, cluster centers are selected randomly and then new cluster centers are allocated according to the Euclidean distance to minimize $J(C)$.
until J(C) almost unchanged. Fig. 5 shows the right hand clustered data by k-means. From Fig. 5, we can see that the original 100 right hand joints are clustered well into 12 classes by which a 100 dimensional vector is formed to represent an arm-hand action.

**BOW Model**

The BOW (i.e., bag of words) model is mainly used for document searching at first which is an orderless model containing the frequency of each word in the article. The importance for a word in an article is proportion to the frequency in this article and in inverse proportion of the frequency in all articles. Usually, the following two weight factors are used:

\[
\text{tf}(\text{term frequency}): \quad \text{tf}_{w,d} = \frac{n_w}{\sum_j n_j} \tag{7}
\]

\[
\text{idf} (\text{inverse document frequency}): \quad \text{idf}_{w,d} = \log \frac{|D|}{|\{d: w \in d\}|} \tag{8}
\]

where, \(n_w\) is the frequency of word in the document \(d\) and \(\sum_j n_j\) is the number of words in the document; \(|D|\) is the number of all documents, \(|\{d: w \in d\}|\) is the number of documents containing word \(w\).

In this paper, BOW model is used to establish the visual equivalent words and the visual codebook is then formed with given cluster centers by counting the frequency of each word for every arm-hand action. Weight factors are then used to decrease redundant information of BOW model. Fig. 6 shows the histograms of BOW model for three actions. The figure shows that the bins of actions are decreased after the weight factor procedure. In addition, the bins of the same arm-hand action are quiet similar but different from another although there are some overlap bins.

**SVM Classifier**

Support vector machines (SVMs) are a set of supervised learning methods for classification, regression and outlier detection. In this paper, SVMs are applied for classifying the code book from BOW model. Table 1 shows the performance of SVMs with different kernel function and parameter and using the linear SVMs can achieve high recognition accuracy.
Table 1. SVMs accuracy with different kernel function and parameters for the in front view data.

<table>
<thead>
<tr>
<th>Kernel type</th>
<th>Parameter</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polynomial</td>
<td>d=3</td>
<td>94</td>
</tr>
<tr>
<td>RBF</td>
<td>C=1.2, g=0.01</td>
<td>86.7</td>
</tr>
<tr>
<td></td>
<td>C=1.2, g=0.005</td>
<td>91.3</td>
</tr>
<tr>
<td></td>
<td>C=1.2, g=0.001</td>
<td>95.3</td>
</tr>
<tr>
<td>Linear</td>
<td>/</td>
<td>96.7</td>
</tr>
</tbody>
</table>

Table 4. Classification Accuracy Comparison with different method on self-built dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuzzy C-means + HMM[6]</td>
<td>82.6</td>
</tr>
<tr>
<td>Covariance Descriptors[10]</td>
<td>92.6</td>
</tr>
<tr>
<td>BOWs + KNN</td>
<td>95.3</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>96.7</td>
</tr>
</tbody>
</table>

Experiments

In order to testify the efficiency of the proposed method, we built a dataset referring to the public dataset MSR Action 3D [6] including 5 arm-hand actions.

Dataset

The self-built dataset contains five arm-hand actions for training and testing: high arm wave, draw circle, hand clap, forward punch and high throw of in front view and 45 degree angle view respectively. These five representative arm-hand actions are chosen because they have motion in different directions, especially in left-right direction and front-rear direction. Fig. 7 shows 10 frames of 5 arm-hand actions and each line presents for one kind of arm-hand action. Besides, each arm-hand action in front view and the 45 degree angle is performed by 10 and 5 person respectively for 3 times. Each time, 100 frames of 20 skeleton joints data are recorded with Kinect sensor.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>96.7</td>
</tr>
<tr>
<td>S2</td>
<td>95.3</td>
</tr>
</tbody>
</table>

Our Proposed Method on Self-built Dataset for in Front View and 45 Degree View

For group S1, k-means is applied directly to the normalized skeleton joint data. During the experiment, we found that 12 cluster centers can help us achieve high recognition accuracy of
arm-hand action. After clustering, 12 cluster centers for all data can be recorded as the words in BOW model. Adjacent right hand joints are clustered into same center. In Fig. 6, we can see that the arm-hand actions from the same category share the similar distribution and some peaks of histograms appears in different arm-hand action which may leads to the misrecognition. For group S2, rotation is necessary before the k-means. The following processes are the same as those of S1. In order to compare the effect of different characteristics of the algorithm, Leave-One-Out method (LOO) is used to evaluate the experiment result. In our experiment, we used all arm-hand action instances from 29 subjects for training and remaining 1 subject for testing. The experiments are conducted for 30 times, excluding one subject in each run. Using this method allows us employ as much data as possible for training. At the same time, it can help us find problematic subjects and analyze the reason of the classification errors.

Table 2 and Table 3 show the confusion matrix of in front view and 45 degree view arm-hand action recognition. The average accuracy is 96.7% and 94.6% respectively. From Table 1, we can see that some arm-hand actions are confused, for example, high throw, high arm wave and forward punch. The reason is that these three actions shared some common features which can be implied in Fig. 6. They have some overlap bins. The other reason may in the consequence of different expression performed by different people for the same action. In this paper, only right hand joint is selected to be trained as the feature. Instead, more joints can be used to develop some new features.

**Comparative Methods on Self-Built Dataset for in Front View**

Particularly, we applied some other action recognition methods based on 3D skeleton joints on our self-built dataset. Table 4 shows the recognition accuracy of our proposed method compared with some other action recognition method.

Via fuzzy k-means and HMM [6] can reach 82.6% accuracy. The reason why the recognition accuracy is not high may because of some outliers in training data causing the parameter problems. Besides, the data preprocessing is very important to the HMM application and it needs plenty of time to train the model. The combination method of covariance descriptors and linear SVMs [10] receives 92.6% accuracy. Despite the simplicity of the covariance descriptors, the feature can still represent the actions well. Our approach achieves 96.7% classification rate which is a little higher than KNN method. But it is worth noting that KNN needs much more time to classify because it has no training time and the error cannot be controlled as SVMs. So it is suitable for small dataset. The proposed method in this paper uses k-means algorithm to cluster the skeleton joints into 12 sorts. That can be considered each arm-hand action is resolved into 12 sub-actions. The same kind of arm-hand action has similar sub-actions, and the application of BOW model calculates the frequency of every sub-action which can represent the feature of these arm-hand actions well. It is simple in calculation and easy to understand the physical meanings.

**Conclusions**

In this paper, a combination of BOW model and SVMs framework is proposed for arm-hand action recognition. By training different number of cluster centers for k-means and selecting kernel functions of SVMs for each action, the BOW model and linear SVMs framework can work efficiently. The feature of BOW model based on k-means algorithm using joint location we introduced in this paper can represent each arm-hand action well. We evaluated the proposed method on our self-built dataset for in front view and 45 degree view respectively which achieves up to 96.7% and 94.6%. The effectiveness of our approach is forecasted by simplicity and compared to other methods, which reveals its practical advantage. For some simple arm-hand actions in different views regardless of temporality, the simulation experiments show that our proposed method outperforms the state of art in arm-hand action recognition.

However, the proposed method does not capture the order of arm-hand action in time. Therefore, if the given frames are randomly shuffled, the codebook of BOW model will not change. So the proposed method cannot be employed to some reverse temporal order arm-hand action, for example. For this problem, it remains further studies.
References


