A Recognition Method of Traffic Directing Gesture based on Multi-feature Extraction and Sparse Coding

Xuan WANG*, Lanxiang CHONG

College of Information, Northwest University, Xi’an 710061, China

Abstract

According to the characteristics of traffic directing gesture, this paper presents a multi-feature extraction method based on spatio-temporal motion distribution and trajectory features of gesture. This method exploits sparse representation to reduce dimensionality of feature vectors and then achieves classification and recognition for traffic directing gesture based on the Locality constrained Linear Coding and the K-nearest neighbor classifier. The experiment results prove the accuracy and effectiveness of this proposed algorithm.

Keywords: Traffic Directing Gesture; Multi-feature Extraction; Motion Frequency Image; Online Dictionary Learning; Locality-constrained Linear Coding

1 Introduction

With the development of artificial intelligence, pilotless automobile has started to experiment on the highway. Pilotless automobile technology is based on laser ranging, radar, machine vision and GPS, which only suitable for ideal conditions such as wide road, fine weather and normative traffic signs [1]. When there are no traffic signs or encounter road construction or traffic accident, AI system will not know what to do because the traffic gesture directed by the traffic police could not be recognized.

Along with the process of industrialization, China has become an automobile manufacture and consumption country [2]. Traffic jam has become a serious social problem because of the increasing number of vehicles. It is longer enough to only rely on the stationary mechanical traffic signs, especially in the morning and evening peak hours or heavy traffic roads. So it is necessary that traffic police use gesture to direct traffic flow at crowded crossroads.

Traffic directing gesture recognition is a difficult theory in unmanned and auxiliary driving technology. It can be classified into two categories, technology based on motion capture or machine vision. The motion capture based technology has high precision and efficiency, but it is expensive and complex which needs a series of equipments such as sensors, signal acquisition...
device, data transmitting and processing equipments. The machine vision and image processing based technology is getting much more important in recent years because it is much simple and cheaper in use.

In this paper, we use static camera to capture video stream at crossroads. Then process the video stream and recognize directing gesture through hardware bundled with the camera. Finally transmit directing signals to the pilotless automobile system. Our main research is focus on the methods of feature extraction and recognition.

Traffic directing gesture recognition is a kind of action and behavior recognition technology. In previous researches, the major research methods include optical flow method [3], global spatio-temporal feature method [4], spatio-temporal interesting point method [5] and so on. These methods both have their own advantages and disadvantages. Considering traffic directing gesture is simple, standard and the position concentrating on arms and hands, a precision and real time algorithm combined with these characteristics is necessary.

In this paper, a multi-feature extraction and sparse coding classification method is proposed. The mainly steps of the method are shown as follows:

1. The interval sampling is abstracted in the time axis of captured video stream, which is called time window as the basic unit of features.
2. Extracting features from every time window by the multi-feature extraction algorithm.
3. Set up encoding dictionary by online dictionary learning (ODL), and apply the dictionary for Locality-constrained linear sparse coding (LLC) to reduce dimensionality of each time window feature.
4. Classify the input gesture samples by the K-nearest neighbor classifier.

The block diagram of the proposed method is shown in Fig. 1.

![Fig. 1: Block diagram of the proposed method](image)

The rest of this paper is organized as follows. Section 2 describes the concrete steps of multi-feature extraction algorithm including motion frequency image (MFI), motion history image (MHI) and spatio pyramid matching (SPM) method. Section 3 describes the sparse coding and gesture classification method. The experimental results and analysis are given in Section 4. Section 5 is conclusions.

### 2 Multi-feature Extraction Algorithm

Feature extraction and description algorithm is the main point and difficult point in real time human action recognition system. In designing of the system, a delicate balance must be maintained.
between recognizing accuracy and computing efficiency. To achieve this purpose, generating \( L \)-width time window by every successive \( L \) frames from the origin video stream. And set \( f \) as the sparsity of time window, which signifies the number of overlapping frames between two adjacent time windows is \( f \) (\( 0 \leq f < L \)). Choose time window as the basic unit for feature extraction is necessary for the correlogram feature extraction algorithm.

Correlogram based action feature extraction is a kind of global spatio-temporal feature algorithm. Its principle is to transform the three-dimensional matrix of time window into a two-dimensional correlogram according to some algorithms. This method creates a mapping from three-dimensional data to two-dimensional data, which greatly reduce the computing amount of data then improve efficiency. In the previous researches on correlogram, motion energy image (MEI) which describe the motion intensity distribution and motion history image (MHI) which describe the motion trajectory in time window are widely used. In this paper, a correlogram called motion frequency image (MFI) is proposed. MFI which improved from MEI describes spatial distribution and motion intensity of actions. However MFI doesn’t include motion trajectory features and the difference computing of the inside points of moving object may produce hollow zones and lead to inaccuracies. In order to overcome these shortcomings, we combine MFI with MHI to describe the motion features together for a same time window.

### 2.1 Construction of the global feature vectors by MFI and MHI

To compute MFI, for each frame \( I(x, y, t) \) in an \( L \)-width time window, at first its three frame difference image \( D(x, y, t) \) can be calculated by the following equation:

\[
D(x, y, t) = |I(x, y, t + 1) - I(x, y, t)| + |I(x, y, t) - I(x, y, t - 1)|
\]  

Accumulate difference images of every three adjacent frames in a time window, the gray energy image can be calculated by:

\[
P_L(x, y) = \sum_{t=1}^{L} D(x, y, t)
\]  

Normalize \( P_L(x, y) \), and square it to weaken noise points which have lower frequency. We can get MFI by the following equation:

\[
MFI(x, y) = [P_L(x, y) / \max (P_L(x, y))]^2
\]

Then mapping the values of all pixels in MFI to a histogram, choose non-uniform quantization histogram as the feature vectors to describe motion correlogram. Assuming the non-uniform quantization histogram has \( N \) intervals, the \( i \)-th interval determined by:

\[
\left[\frac{i(i-1)}{N(n+1)}, \frac{i(i+1)}{N(N+1)}\right]
\]

Finally accumulate the number of pixels in each interval of MFI histogram and calculate the proportion of every interval. The generated \( N \)-dimensional histogram vectors of two different MFI images (Stopping and changing lane) are shown in Fig. 2.
MHI [6] can describe the overall trend of motion and the information of motion direction. It can be defined as:

$$MHI(x, y, t) = \begin{cases} L & \text{if } DI(x, y) > T \\ \max(0, MHI(x, y, t-1) - 1) & \text{otherwise} \end{cases}$$ (5)

where $MHI(x, y, t)$ is the motion history image of the $t$-th frame, $T$ is binaryzation threshold of $DI(x, y)$. $DI(x, y)$ is the difference image between two adjacent frames which can be written as:

$$DI(x, y) = I(x, y, t) - I(x, y, t-1)$$ (6)

For each time window, the last frame $MHI(x, y, t)$ is set to be the MHI for the corresponding time window. And then calculate the histograms of oriented gradients (HOG) [7] of MHI. The steps of HOG algorithm is shown as follows:

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2}$$ (7)

$$\theta(x, y) = \arctan(G_x(x, y)/G_y(x, y))$$ (8)

where $G_x(x, y)$ and $G_y(x, y)$ are the oriented gradients of $x$-axis and $y$-axis in MHI. Divide the range of gradient oriented angle $\theta(x, y)$ ($\theta(x, y) \in [-\pi/2, \pi/2]$) into $M$ quantization bins. Accumulating gradient magnitude $G(x, y)$ of all the pixels in each bin, we can get the $M$-dimensional HOG vector from each time window.

Whatever the dimensionality of feature vectors is $M$ and $N$, the small amount of data they included is not enough to describe motion features. So we build ROI templates for a further construction to the specific region of correlograms. The ROI templates are shown in Fig. 3, the white zones of templates is the regions of interest (ROI).

![ROI templates](image)

Fig. 3: The ROI templates we used in this paper

Recalculate histogram vectors for MFI and MHI with above templates, and connect all vectors of all templates as the final feature vector. If the number of templates is $K$, the dimensionalities of total feature vector computed by the above-mentioned algorithm is $K(M + N)$. 
2.2 Construction of the local feature vectors by SPM

The robustness of global spatio-temporal features may be affected by some environment factors, such as body shaking and background noise. But local features extracted from the most strong motion zone in correlogram of the same gesture samples are similar. According to this characteristic, local spatio-temporal feature can be described through the combination of above-mentioned algorithms with the spatial pyramid matching (SPM) method. Combing global features with local features, we can get a high dimensional vector from each time window.

SPM [8] is a method to describe the relationship between global features and local features in spatial structure. According to SPM, image is divided into several sub blocks with different scale in different layer, and then applies feature extraction and sparse coding for all the sub blocks in each layer. In the recognition process, the corresponding sub blocks will be matched respectively.

In this paper, we divide the correlograms into several sub blocks in same scales. Accumulate all the pixel values in every adjacent $4 \times 4$ sub blocks. There is no doubt that the most strong motion zone has the maximum sum among all the $4 \times 4$ zones. Apply three layers SPM to this zone, which divide correlogram into 21 sub blocks in total, as shown in Fig. 4. At last, calculate $K(M + N)$ dimensional vector for each sub block by MFI, MHI and ROI templates. The total dimensionality of feature vector is $21K(M + N)$.

![Fig. 4: Three layers SPM of the MHI of turning left gesture](image)

In conclusion, this multi-feature extraction algorithm combined MFI, MHI and SPM can cover their shortages. For example, MFI describes the spatial distribution of movement in a time window, but it doesn’t include the subsequence and trajectory information. And MHI cannot describe the motion frequency information. Moreover, this algorithm describes the local features in spatial domain by SPM. The experiments prove that this algorithm has certain robustness in the changing of illumination, scale and rotation, and the computing efficiency and time complexity of the algorithm are in the allowed range.

3 Sparse Coding and Classification

After extracted the high dimensional feature vectors, the next step is to find suitable classifier and dimensionality reducing method. In this paper, we propose an action classification method which combined sparse dimensionality reducing method, locality-constrained linear coding (LLC) method and the K-nearest neighbor (KNN) classifier.
3.1 Sparse dimensionality reducing

In order to improve the efficiency of classification, it is necessary to reduce the high dimensional vectors. The fundamental of dimensionality reducing can be represented as follow:

\[ X = D\alpha^T \] (9)

where \( X \in \mathbb{R}^{m \times n} \) is the original sample matrix and \( D \in \mathbb{R}^{m \times k} \) is the generated dictionary matrix. If the linear combination of any atoms in \( D \) can represent \( X \) and \( k \ll m \), the index matrix \( \alpha \in \mathbb{R}^{n \times k} \) can approximate used as the reducing dimensionality of the origin sample matrix \( X \). Therefore we focus on how to generate the dictionary matrix.

Online dictionary learning (ODL) [9] is a kind of dictionary learning method based on least angle regression and sparse representation. If the column vectors in \( X \) can be represented as \([x_1, ..., x_n]\), the solving process of \( D \) is shown as follows:

\[
\alpha_t \triangleq \arg \min_{\alpha} \frac{1}{n} \sum_{i=1}^{n} \left( ||x_i - D_{t-1} \alpha_i||_2^2 + \lambda ||\alpha_i||_1 \right) 
\] (10)

\[
A_t \leftarrow A_{t-1} + \alpha_t \alpha_i^T 
\] (11)

\[
B_t \leftarrow B_{t-1} + x_t \alpha_i^T 
\] (12)

\[
D_t \triangleq \arg \min_{D} \frac{1}{t} \left( ||x_i - D\alpha_i||_2^2 + \lambda ||\alpha_i||_1 \right) 
= \arg \min_{D} \frac{1}{t} \left( \frac{1}{2} Tr(D^TD\alpha_i) - Tr(D^TB_i) \right) 
\] (13)

where \( t \) is the iterations, \( \lambda \) is the standard factor, \( \alpha = [\alpha_1, ..., \alpha_n] \) is the sparseness index matrix. Iterative computing by the above equations for several times then we can get the optimal dictionary \( D \). The dimensionality of reduced feature vector is equal to the scale \( k \) of dictionary \( D \), and the compression rate is \( 21K(M + N)/k \).

3.2 LLC and KNN classifier

Locality-constrained linear coding (LLC) [10] is a sparse coding method for SPM features. It can be defined as follow:

\[
\arg \min_{C} \sum_{i=1}^{n} \left( ||x_i - Dc_i||_2^2 + \lambda ||d_i \cdot c_i||_2^2 \right) 
\] (14)

We can see from Eq. (14) that LLC has a local adapter \( d_i \) for the elements in dictionary, which is different from previous sparse coding methods. According to the distance to each sample vector, the local adapter distributes different degree of freedom for each element as the follow equation:

\[
d_i = \exp \left( \frac{\text{dist}(x_i, D)}{\theta} \right) 
\] (15)

where \( \text{dist}(x_i, D) \) represented calculating Euclidean distance between \( x_i \) and every atom in \( D \). \( \theta \) is the attenuation factor, which regulates the attenuation speed of \( d_i \). In this paper, we choose \( D \) as the dictionary of LLC.
In the coding of spatial pyramid local features, we choose the max pooling [11] method to calculate the max pooling feature for each sub block with different scales. Then connect those features in every scale together to generate the final feature vectors. The max pooling method equation is shown as follow:

$$\beta = \xi_{\text{max}}(\hat{c})$$  \hspace{1cm} (16)

where $\hat{c} \in \mathbb{R}^{k \times n}$ are the sparse codes of local features, and $\xi_{\text{max}}$ means to calculate $\beta$ for each column in $\hat{c}$. For the $i$-th column, the vector $\beta_i$ can be represented as follow:

$$\beta_i = \max (|\hat{c}_{i1}|, |\hat{c}_{i2}|, ..., |\hat{c}_{in}|)$$  \hspace{1cm} (17)

The above-mentioned LLC and local layering features make up the sparse global and local features. The weight of the local feature is needed before classifying through KNN. The following equation is used to set weights for the SPM local layering features with different scales:

$$\beta_l = \frac{1}{2^{l-1}} \beta$$  \hspace{1cm} (18)

where $l$ is the $l$-th layer in SPM, and $\beta_l$ is the weight of the sub block features in the $l$-th layer.

Apply the K-nearest neighbor (KNN) classifier to the input captured gesture features to be detected. The mainly steps are shown as follows:

1. Calculate the Euclidean distance between the input features and the sample features with same gesture label.
2. The time window feature vector is regarded as the basic unit of samples, and all the time window features in same gesture sample is regarded as the sample feature pool.
3. Voting to the gesture label which has the smallest Euclidean distance.
4. Select the gesture label which have the most amount of votes as the recognize result.

4 Experiment Results and Analysis

This section presents the experimental results of the recognition system we designed. In order to choose the test samples, we record gesture video fragments recording by a fix camera in a crossroad. The resolution of the video is $320 \times 240$, and the format is AVI. As shown in Fig. 5, there are five common traffic directing gestures we tested include stopping, going straight, turning left, change lane and slowing down.

![Fig. 5: Test samples of traffic directing gestures](image)

According to a large number of experiments, the appropriate parameters in the method are shown in Table 1.
Table 1: Values of parameters in this paper

<table>
<thead>
<tr>
<th>N</th>
<th>M</th>
<th>L</th>
<th>f</th>
<th>K</th>
<th>fps</th>
<th>k</th>
<th>t</th>
<th>λ</th>
<th>θ</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>8</td>
<td>25</td>
<td>5</td>
<td>15</td>
<td>15</td>
<td>250</td>
<td>100</td>
<td>500</td>
<td>100</td>
</tr>
</tbody>
</table>

We can see from Table 1 that the dimensionality of the final feature vectors is \(21K(M+N) = 5040\), and the reduced dimensionality is \(k = 250\).

The detected results of the test sample are shown in Table 2. That shows this method have high recognize rates for the traffic directing gestures which have long durations and circulate positions, and low recognize rates for the gestures which have short durations and simple positions.

Table 2: Recognition rate of the test samples

<table>
<thead>
<tr>
<th>Gesture label</th>
<th>The number of test samples</th>
<th>The number of correctly tested samples</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Going straight</td>
<td>18</td>
<td>16</td>
<td>88.9%</td>
</tr>
<tr>
<td>Stopping</td>
<td>20</td>
<td>14</td>
<td>70.0%</td>
</tr>
<tr>
<td>Turning left</td>
<td>20</td>
<td>19</td>
<td>95.0%</td>
</tr>
<tr>
<td>Changing lane</td>
<td>15</td>
<td>12</td>
<td>80.0%</td>
</tr>
<tr>
<td>Slowing down</td>
<td>15</td>
<td>11</td>
<td>73.3%</td>
</tr>
</tbody>
</table>

In order to test real time capability of the system, we measure the average time of computing for every time window. The average time results are shown in Table 3. The code is implemented in MATLAB, and the evaluation is carried out on a PC with 2.8G CPU and 2GB ram.

Table 3: Time consuming of algorithm steps

<table>
<thead>
<tr>
<th>The main steps for single time window</th>
<th>Average time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-feature extraction</td>
<td>0.1869</td>
</tr>
<tr>
<td>Sparse coding</td>
<td>0.0309</td>
</tr>
<tr>
<td>KNN classifier</td>
<td>0.0723</td>
</tr>
<tr>
<td>Total</td>
<td>0.2901</td>
</tr>
</tbody>
</table>

As Table 3 shows, when the sparsity of time window \(f = 5\), the theoretical frames per second (fps) is 17.2354. Therefore we set the input video frame rate is 15. These results show that the method can adapt to the real time needed when the frame rate is not too high.

5 Conclusion

This paper designs a fast and effective traffic directing gesture recognition system based on multi-feature extraction and sparse coding, which can be used in the pilotless automobile system. The feature extraction method combined MFI, MHI and SPM have good robustness for the changing of illumination, scale and rotation. Then the sparse dimensionality reduction, LLC and KNN
classifier are used in gesture classify. The experiment results show the effectiveness of the system. But the results also show that the method has lower recognition rate for the simple and short actions. How to improve the shortages of the method will be the next problem to be solved.

References


