Journal of System Simulation

Volume 27 | Issue 8 Article 19

8-3-2020

Human Action Recognition Based on Self-Learning Feature and HMM

Jiarui Wen
Shanghai University, School of Mechatronic Engineering and Automation,Shanghai Key Laboratory of Power Station Automation, Shanghai 200072, China;

Lina Liu
Shanghai University, School of Mechatronic Engineering and Automation,Shanghai Key Laboratory of Power Station Automation, Shanghai 200072, China;

Rui Ling
Shanghai University, School of Mechatronic Engineering and Automation,Shanghai Key Laboratory of Power Station Automation, Shanghai 200072, China;

Shiwei Ma
Shanghai University, School of Mechatronic Engineering and Automation,Shanghai Key Laboratory of Power Station Automation, Shanghai 200072, China;

Human Action Recognition Based on Self-Learning Feature and HMM

Abstract

Abstract: The current methods of human action recognition by computer vision are mostly based on hand-craft features and usually prior knowledge-required. They inevitably depend on specific applications and neglect the inner structure of visional information. *A novel method which integrated self-learned pose features and combined posture symbol rules was proposed to achieve the recognition of human action more efficiently*. The structural features of posture silhouette were extracted and a codebook of primary posture was built through the establishment of a sparse auto-encoder network. Then, in the phase of recognition, the Hidden Markov Model was employed to train the models for different action categories. Besides, a key frame extraction algorithm was developed to reduce the redundancy of long code sequence before training HMMs. Simulation experiments manifest the effectiveness of the proposed method.

Keywords

pose recognition, self-learned feature, SAE, HMM, posture codebook

Recommended Citation

Wen Jiarui, Liu Lina, Rui Ling, Ma Shiwei. Human Action Recognition Based on Self-Learning Feature and HMM[J]. Journal of System Simulation, 2015, 27(8): 1782-1789.

第 27 卷第 8 期 系统仿真学报© Vol. 27 No. 8
2015 年 8 月 Journal of System Simulation Aug., 2015

Human Action Recognition Based on Self-Learning Feature and HMM

Wen Jiarui, Liu Lina, Rui Ling, Ma Shiwei

(Shanghai University, School of Mechatronic Engineering and Automation,
Shanghai Key Laboratory of Power Station Automation, Shanghai 200072, China)

Abstract: The current methods of human action recognition by computer vision are mostly based on hand-craft features and usually prior knowledge-required. They inevitably depend on specific applications and neglect the inner structure of visional information. *A novel method which integrated self-learned pose features and combined posture symbol rules was proposed to achieve the recognition of human action more efficiently*. The structural features of posture silhouette were extracted and a codebook of primary posture was built through the establishment of a sparse auto-encoder network. Then, in the phase of recognition, the Hidden Markov Model was employed to train the models for different action categories. Besides, a key frame extraction algorithm was developed to reduce the redundancy of long code sequence before training HMMs. Simulation experiments manifest the effectiveness of the proposed method.

Keywords: pose recognition; self-learned feature; SAE; HMM; posture codebook

基于自学习特征与 HMM 的人体动作识别

温加睿，刘丽娜，芮玲，马世伟

(上海大学机电工程与自动化学院上海市电站自动化重点实验室，上海 200072)

摘要：利用机器视觉进行人体动作识别的方法大多数基于手工特征并需要先验知识，这类方法不可避免地依赖于特定问题而忽略了视觉信息的内在结构。*提出了一种利用自学习特征及姿态组合规则进行有效动作识别的新方法*。使用稀疏自编码(SAE)网络提取轮廓图像的结构特征并构造姿态码本。在识别阶段，使用隐马可夫模型(HMM)训练不同动作类别的模型。设计了一种关键帧提取算法用于在训练 HMM 前降低长序列的冗余度。通过仿真实验验证了该方法的有效性。

关键词：动作识别；自学习特征；SAE；HMM；姿态码本

中图分类号：TP391.1 文献标识码：A 文章编号：1004-731X (2015) 08-1782-09

Introduction

Human action recognition is currently one of the most active research areas of computer vision. It has a very widely practical value in many applications, such as virtual reality, security surveillance, medical monitoring and video analysis. Since there is no universal structure and perfect-defined model for action recognition, the research of this area is significant and full of challenges. Many researches[1-8] on the methodology about building human model and recognizing action have been proposed since the past decade. Generally, human action can be considered as a temporal

Receive: 2015-06-15 Revised：2015-06-24；
Biography：Wen Jiarui (1989-), Male, Guangdong, China, Master, Research Filed: Pattern Recognition, Image Processing, Behavior Recognition; Liu Lina (1981-)，Female, Shandong, China, Doctor, University Lecturer, Research Filed: Pattern Recognition, Image Processing, Face Recognition

variation model[3], whose recognition method can be separated into two major parts: the extraction of human model-based features and Hidden Markov Model.

Human model-based feature extraction means the estimation or reconstruction of human posture, trying to find the representation of the body pose from a single image in monocular vision. In [4-5], a Hu-moment is used to describe the shape of human silhouette. By extracting seven invariant moments, the shape of a posture can be represented as a seven-dimensional feature which has the quality of scale and position invariance. However, the amount of information of this feature is limited and the values of feature are sensitive to occlusion. Some approaches like [2,6] use radon transform or its improved version, called R transform, which also have geometrical invariability like Hu's but has more information, to describe posture image. The domain integrations in 180 degrees give it more ability to represent the structure if an image. Other model-based methods, such as fitting limbs and torso with pictorial structure [7], using curve to extract the template body contours [3] or modeling human body with skeleton vector[4,8] are high algorithm complexity or still ignored inner structure in an image and exist some instability in a certain degrees.

In recent years, there has been a growing research on the self-learning of feature (also called deep learning)[9-10]. Unlike those traditional hand-craft features mentioned above, these kinds of models can automatically construct high-level expression from raw input, and consequently discover their feature. Sparse Autoencoder(SAE) [9] is a kind of model that can learn feature by itself. By training a multilayer neural network whose output layer is approximate to input layer, SAE will learn to encode a high dimensional original data to a low dimensional expression as a new feature. This technique has been used to resolve many recognition problems such as classifying handwritten digits [11], cancer cells [12], and detecting moving body [13]. By applying SAE, a self-learning model can be developed to find the feature of human posture.

Hidden Markov Model (HMM) is a statistical model which has been successfully utilized in long sequence template matching like speech recognition. HMM has shown its well performance in action recognition. Besides, many researchers have developed lots of improved version of HMM [3] to achieve more adaptation in the task of action recognition application.

In this paper, an approach based on HMM using self-learned feature extracted with SAE is developed to solve the human action recognition task. In the proposed method, time-sequential human action images are transformed to a feature vector sequence by extracting a feature from each image using SAE. Each feature vector of the sequence is then assigned a symbol corresponding to a codeword symbol in codebook with softmax classifier [9]. After the assignment, the time-sequential images are converted to a corresponding symbol sequence. During the training phase, an algorithm of key frame extraction is firstly applied to reduce the redundancy of long symbol sequence. Then, the HMM parameters of each category are optimized so as to best describe the training symbol sequences of the same categories of human action. For human action recognition, the model which best matched the observed symbol sequence will be finally selected as the recognized category.

第 27 卷第 8 期 系统仿真学报 Vol. 27 No. 8
2015 年 8 月 Journal of System Simulation Aug., 2015

1 Sparse Autoencoder (SAE)

Sparse Auto-Encoder is a kind of multi-layer feed-forward neural network which is able to learn unsupervised feature automatically and can be trained with back- propagation algorithm [12]. By learning an approximation to an identity function, SAE tries to find the correlation among the data and to construct a better feature representation from high-dimensional input.

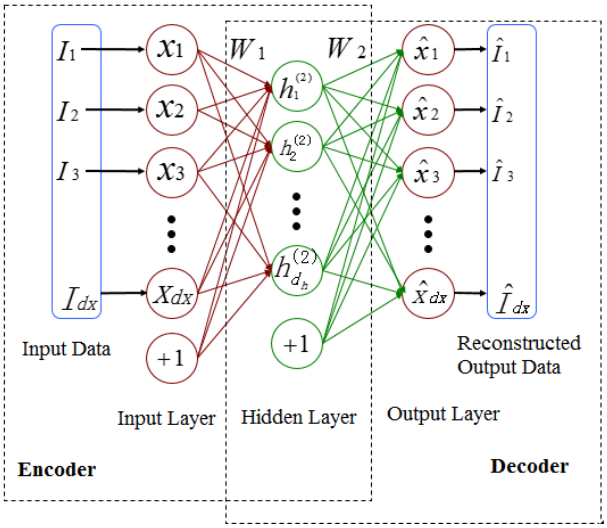

Fig.1 The architecture of 3-layer Sparse Autoencoder

As shown in Figure 1, the neural network consists an encoder and a decoder, determined by the parameters weights and biases. The encoder in the input layer transforms input data x to a corresponding activation h, which can be seen as a new compressed feature representation of x. The decoder in output layer is trained to reconstruct an approximation $\hat{x}$ of input from representation h.

Therefore, the training of an autoencoder is meant to narrow the discrepancy between x and $\hat{x}$ by optimizing weights W and biases b.

The entire training set is denoted as:

$$X = (x^{(1)}, x^{(2)}, ..., x^{(m)})^T$$

where $x^{(k)} \in \Re^{d_x}$. Here, m and d_x are the number of samples training set and the dimension of each vector (the number of pixels in each image I), respectively.

$$h^{(l)}(k) = (h_1^{(l)}(k), h_2^{(l)}(k), ..., h_{d_h}^{(l)}(k))^T$$

is the feature learned from the k-th input at layer l, where d_h is the number of hidden units in layer l. The discrepancy between x and $\hat{x}$ can be described with the as following cost function:

$$J_{SAE}(W,b) = \frac{1}{m}\sum_{k=1}^{m}\left(\left\|H_{w,b}(x^{(k)}) - x^{(k)}\right\|^2\right) +$$

$$\frac{\lambda}{2}\|W\|_2^2 + \beta\sum_{j=1}^{d_h}\left[\rho\log\frac{\rho}{\rho_j} + (1-\rho)\log\frac{1-\rho}{1-\rho_j}\right] \quad (1)$$

The first term of the cost function is a traditional squared error which describes the average discrepancy of input $x^{(k)}$ and its reconstruction $\hat{x}^{(k)} = H_{w,b}(x^{(k)})$ over the entire dataset. Here, $H_{w,b}(x)$ is an encode-decode procedure:

$$H_{w,b}(x) = d_{w_2,b_2}(e_{w_1,b_1}(x))$$

The encoder $e = s(W_1 x + b_1)$ takes $x \in \Re^{d_x}$ as input and maps it to a new representation $h \in \Re^{d_h}$ in hidden layer, where s is a non-linearity activation function such as sigmoid function:

$$S(z) = (1+\exp(-z))^{-1}$$

The decoder $d = s(W_2 x + b_2) = \hat{x}$ then maps e back to a reconstruction d, a prediction of x, given code value e. Here, $W_1 = W \in \Re^{d_h \times d_x}$ is a weights matrix. To half reduce the number of weights matrix parameters, the tied weights method is used, the weights W_2 is set equal to W^T. $b_1 \in R^{d_h}$ and $b_2 \in R^{d_x}$ are the biases in input layer and hidden layer, respectively.

The second term is a regularization term with a weight decay parameter λ. This term tends to decrease the magnitude of the weight, and helps prevent over-fitting. It can be calculated in:

$$\| W \|_2^2 = \sum_{l=1}^{n_l} \sum_{i=1}^{S_l - 1} \sum_{j=1}^{S_l} (w_{i,j}^{(l)})^2 = tr(W^T W) \quad (2)$$

Here, n_l is the number of network layers and s_l is the number of units in the l-th layer. The parameter $w_{i,j}^{(l)}$ represents the connection weight between the i-th neuron in layer $l-1$ and the j-th neuron in layer l.

The third term can be written as

$$\beta \sum_{j=1}^{S_2} KL(\rho \| \hat{\rho})$$

This item is added to achieve sparsity. Here, $KL(\rho \| \hat{\rho})$ is the Kullback-Leibler(KL) divergence between average activation $\hat{\rho}_j$ of hidden unit j, and its desired activations ρ_j, which is a small value close to zero. In this paper, the model parameters of SAE are $d_x = 900$, $d_h = 200$, $n_l = 2$.

2 Hidden Markov Model (HMM)

HMM is a kind of stochastic state transit model. It treat discrete time sequences as the output of a Markov process whose states cannot be directly observed. A discrete HMM can be represented as $\lambda = (N, M, A, B, \pi)$, where N is the number of hidden states:

$$Q = \{q_1, q_2, \cdots, q_N\}$$

while M is the number of observable output symbols:

$$V = \{v_1, v_2, ..., v_M\}.$$

For others parameters, $A = [a_{ij}]_{N \times N}$ is a matrix of state transition probabilities, where $a_{ij} = P(s_{t+1} = q_j \mid s_t = q_i)$ is the probability of transiting from state q_i to state q_j. $B = [b_j(k)]_{N \times M}$ is a matrix of observation probabilities, where $b_j(k) = P(v_k \mid s_t = q_j)$ is the probability of observing symbol v_k at state q_j. π is initial probability distribution, here $\pi = \{\pi_i \mid \pi_i = P(s_1 = q_i)\}$.

A simple example of semi-connected HMM, where $N = 5$ and $M = 3$ is shown in Figure 2. The length of observing time is T.

Recognition base on HMMs involves two parts: the first part is to train the HMMs for each categories and the second part is to calculate each model's probability of observation O output by the model λ_i, which can be written as $P(\lambda_i \mid O)$. Given a symbol pattern, the model which corresponding to the maximum $P(\lambda_i \mid O)$ is the most likely model to predict. We can use Baum-Welch algorithm [14] to optimize λ and use the forward or backward algorithm [14] to calculate $P(\lambda \mid O)$.

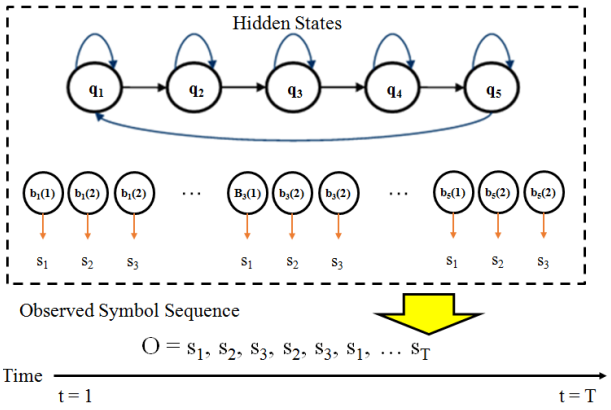

Fig. 2 A Semi-Connected HMM Cycling from Left to Right

3 Proposed Method

3.1 System Overview

The architecture of the system in this paper consists two major parts, including a feature-learning codebook framework and an action recognition module.

For the codebook framework, SAE is trained to learn the features of human postures, using unlabeled silhouette images as training set. After training of SAE, the hidden layer, is connected to a Softmax classifier and use labeled primary posture image set to optimize the parameters of this new architecture again. Finally, this optimized architecture (See Figure 3) will become the posture codebook, which will be used to match a posture image to its corresponding

第 27 卷第 8 期 系统仿真学报 Vol. 27 No. 8
2015 年 8 月 Journal of System Simulation Aug., 2015

symbol. Therefore, an action sequence with a range of posture images will be transformed to a symbol sequence.

The action recognition module involves two phases: training and recognition. In training part, a key frames extraction algorithm is designed to reduce the redundancy and use HMM to describe the models of different actions by training the optimized parameters of their models. By computing the probability of all action categories and the selection of maximum probability, the recognition is achieved. An unknown action symbol sequence will be recognized to the category which is corresponding to the maximum probability.

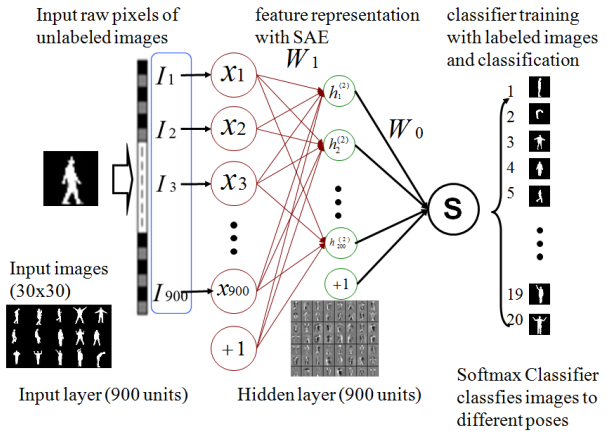

Fig.3 The architecture of SAE and Softmax for primary pose classification

3.2 Feature Learning and Symbol Matching

The first part of the pose recognition is to build the posture codebook, which extracts the self-learned feature from images and match them to different symbols. The self-learned feature is extracted from the normalized images dataset using sparse auto encoder. Then the feature will be classified to different primary posture symbols with Softmax regression. The key postures of different actions are chosen empirically.

To train the whole SAE framework (including a basic SAE and Softmax classifier, see Figure 3), the greedy layer-wise algorithm [10] is employed to trains each layer in turn (In this approach, the SAE has one layer). It learn the feature of the corresponding input image x. Following this, the hidden layer (feature layer) should be connected to a Softmax classifier, then train this new architecture again. Finally, the input layer, hidden layer and Softmax classify layer are combined together to form a codebook capable of predict the input silhouette to digit labels.

In this approach, it use 5000 unlabeled images to train the SAE, and empirically label another 200 images to 20 different primary poses to train Softmax. Therefore the scale of codebook is 20, which means an unknown action images sequence can be represented using 20 different words.

3.3 Action Recognition

Human behaviors are usually kinds of periodical time sequences[4], whose current states are determined by their last states (see Figure 4, it shows a periodical process of waving one hand). Therefore, a Markov process can be used to represent an action and use HMM to recognize disperse symbol sequence of different categories.

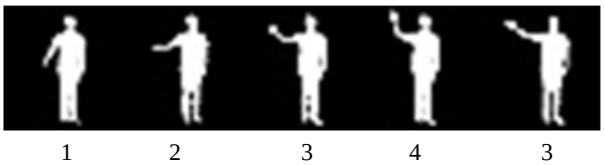

Fig.4 Action "one hand-wave (wave1)" is a Markov process

3.3.1 Key Frames Extraction

After the match of action image sequences, HMM can be applied to learn the model of each action. However, because the action frames of each sample haven't constant length, training HMM is difficult. Therefore, it is important to apply a key

frames extraction before training HMM. Considering human behavior is a periodic process, a piecewise linear compress algorithm can be used to shorten the whole input vector to a constant length but try to keep the periodicity in each sequence.

The key frames extraction algorithm listed below and the length of key sequence is defined as k :

(1) Calculate the length of input sequence L ;

(2) For each piece of sequence data with same value, mark their length as $L(i)$;

(3) Extract the top-$l(i)$ of each piece ordering to a new sequence, where $l(i) = ceil(L(i) \times 10 / L)$;

(4) The top-k codes from the new sequence is the key frames sequence vector V .

Figure 6 shows a side walking movie which has too much redundancy frames between key postures of a whole sequence. The length of key sequence k is set equal to ten. After applying key postures extraction (See Figure 5 and Figure 6), the original image sequence has turn to a shorter one. It can be seen that the extracting algorithm try to keep the action periods as more as possible and maintain the posture frequency between each action period.

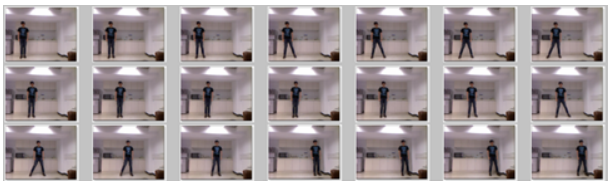

Fig.5 a Long Action Movie

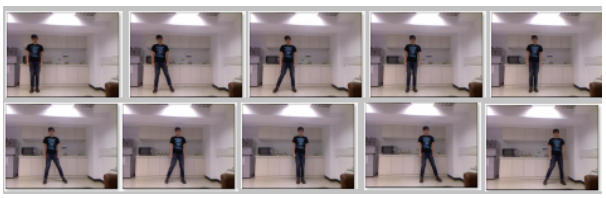

Fig.6 the 10 Key Frame Extracted form Long Images Sequence

3.3.2 Action Recognition

The action models are denoted with λ_i , where i represent the category label of the i-th action, and transform the pose images to codes with the codebook mentioned above. After training each model, $P(\lambda_i \mid O)$ can be calculated for each action model, given an observation posture sequence O . Then the category of sequence O will be recognized as: Category of $O = \arg\max_i P(\lambda_i \mid O)$.

4 Dataset and Experiments

In this approach, a public dataset of Depth Human Action (DHA) [15] is used to train and verify our method. DHA dataset contains the depth information and their corresponding silhouette images of many action, every actions are performed by 21 different actors. We choose six common actions: bend, jack, sidewalk, walk, one-hand-wave (wave1), two-hands-wave (wave2) to conduct the experiment. To evaluate the SAE feature, a comparison is launched in the feature codebooks performance between SAE feature and three conventional hand-craft features: Hu moment (denoted as Hu) [4], Radon transform(denoted as R) [2] and star-skeleton(denoted as Sk) [3]) and use the same framework of HMM for action recognition.

4.1 Codebook Building

We randomly choose two actors to build the codebook for different features and 20 primary postures are empirically labeled from 200 pictures to form the unified training set, which will be applied to extracted four kinds of features and their corresponding codebooks. To decrease the scale of the images, a square object-centered ROI is firstly applied to detect the human silhouette region and zoom out the ROI size to 30x30 pixels. Matched with

第 27 卷第 8 期 系统仿真学报 Vol. 27 No. 8
2015 年 8 月 Journal of System Simulation Aug., 2015

different feature codebooks, all the postures sequence of the other 19 actors are transformed into four kinds of symbol lists. After applying key frames extraction, the symbol lists can be used to train HMMs.

4.2 Model Parameters and Evaluation

In this approach, a five states semi-connected HMM [2] (see Figure 7) is applied and the initial parameters of HMM are set as follow:

(1) The number of hidden states $N = 5$

(2) The number of observable values $B = 20$

(3) The initial values of states transition probabilities

$$A: \begin{bmatrix} 0.5, 0.5, 0.0, 0.0, 0.0 \\ 0.0, 0.5, 0.5, 0.0, 0.0 \\ 0.0, 0.0, 0.5, 0.5, 0.0 \\ 0.0, 0.0, 0.0, 0.0, 0.5 \\ 0.5, 0.0, 0.0, 0.0, 0.5 \end{bmatrix}$$

(4) The initial values of probabilities matrix:

$B = (5\times20) = 1/20$

(5) The initial probability matrix $\pi = [1,0,0,0,0]$

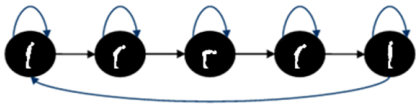

Fig.7 State transition between posture segments

With feature matching and extracting key frame, the entire original image dataset is become a set of digit symbol sequence which can be directly applied to the final training and testing of HMMs. In order to compare the effect of different characteristics of the algorithm, Leave-One-Out method (LOO) is used to evaluate the experiment result.

4.3 Experiments Result

All the self-learned features obtained by SAE are shown in Figure 8, and in order to be more clearly to see, a part of the images enlarge. It shows that the SAE successfully learn the motion of body parts in a certain extent. By observing the brighter areas of the figure, it is easy to find that the SAE has discovered some interesting structure between posture images. Intuitively, some parts of feature reflect the components of human body, such as head, arms, and legs, and the other parts reflect the move direction of limbs and torso.

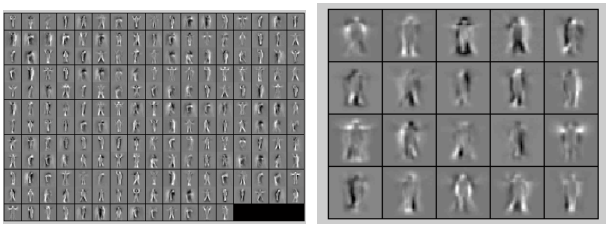

Fig.8 The Visualize Feature learned by SAE

Table 1 illustrates the confusion matrix of the verifying of SAE+HMM model. The data in the table record the average detect number of 10 LOO procedure. In the table, the value of category bend, walk, wave1 are 21, which indicates these classes of action can be accurately identified. For the misjudging category, two sample in sidewalk are predicted to a relatively similar categories walk, one sample in sidewalk is recognized as wave2, one sample in jack is mistaken as wave2.

Table 1 The result of model verifying using Leave-One-Out

	Detection(n/10)					
	bend	jack	sidewalk	walk	wave1	wave2
bend	21					
jack		20				1
sidewalk			18	2		1
walk				21		
wave1					21	
wave2		1				20

After analysis, the results of mistaken are mainly because the key postures of different actions are similar so that the difference of a single image feature may be very small. Moreover Some actors in

dataset perform some extreme posture. Such abnormal condition cause the error when codebook matching. This is also the shortage of label small training set by empiricism which could not take many extreme condition in consideration. These condition can be improved by enlarging the training set of more different condition and adding additional judgment by detecting the moving state of human body.

Figure 9 shows the comparison of the recognition accuracy of four different models, based on the four features mentioned, also use LOO method for verifying. In this histogram, the labels on X-axis are corresponding to different action, Y-axis is the average accuracy. The average of the accuracy of different methods are: SAE method = 96.0%, Skeleton method = 89.5%, Hu method = 80.8%, R method = 89.7%. It can be seen that, comparing to other three hand-craft features, SAE features performs a better recognition result on average.

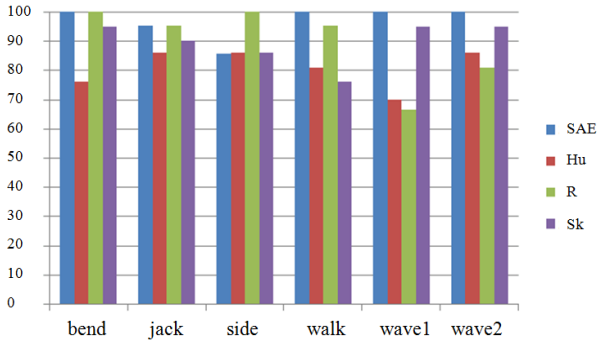

Fig. 9 Accuracy of different methods

The experimental results show that the SAE have a well performance on learning a better description of the image and have a good ability to identify the action by combining SAE feature codebook and HMM. All experiments are carried on the PC (Intel Core (i5) 2.67GHz processor with 4 GB of RAM). The software implementation was performed using Python + MATLAB.

5 Conclusion Remarks

In this paper, a self-learned human action recognition system based on SAE codebook and HMM is developed. For one thing, the framework of SAE codebook successfully extracts the self-learned features which present the vibration and the structure of posture images. For another, the key frame extraction algorithm can effectively reduce the redundancy of primary symbol sequence before training HMMs. By given a primary posture sequence, the HMMs of different actions categories are trained to well recognize actions,. The simulation experiment results show that the SAE+HMM based classifier outperformed some conventional image features based classifiers.

Reference:

[1] GU Jun2xia, DING Xiao2qing, WANG Sheng2jin A Survey of Activity Analysis Algorithms [J]. Journal of Image and Graphics(S2301-3699), 2009, 14(3): 377-387.

[2] Liang Penghua. Research of Human Action Recognition Based on HMM[D]. China: Lanzhou University, 2012.

[3] YU Hui, GUO Li. A Posture Description based on Star-skeleton and HMM[J]. Communications Technology (S1002-0802), 2012, 45(12): 91-94.

[4] Li-li Zheng. the Study of Abnormal Human Behavior Detection Algorithm Based on SVM[D]. China: Zhejiang University of Technology, 2012.

[5] Ahad MAR, Ogata T, Tan JK, *et al*. Moment-based human motion recognition from the representation of DMHI templates[C] //USA: Sice Conference. IEEE, 2008:578 - 583.

[6] Zhang Xu-dong, *et al*. Human Activity Recognition Using Multi-layered Motion History Images with Time-Of-Fligh (TOF) Camera[J]. Journal of Electronics and Information Technology(S1064-2269, 2014, 36(5): 1139-1144.

[7] Tian J, Li L, Liu W. Multi-Scale Human Pose Tracking in 2D Monocular Images[J]. Journal of Computer & Communications(S2027-5227), 2014, 2(2): 78-84.

（下转第 1795 页）

