International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 8, No. 1 (2015), pp. 23-32
http://dx.doi.org/10.14257/ijsip.2015.8.1.03

Human Behavior Recognition based on Conditional Random Field and Bag-Of-Visual-Words Semantic Model

Fengju Bu

School Information Science and Engineering, Shandong Agriculture and Engineering University;
Jinan, Shandong 250100, China
bufengju@126.com

Abstract

Although current gymnastics action detection algorithm has good detection and recognition results but cannot effectively identify a variety of consecutive gymnastic actions and many gymnastics has high false rate. So on this paper we improve the CRF model and bag-of-visual-words semantic model, combine the advantages of both models to build a hierarchical model for behavior recognition, first we create a hierarchical semantic mark CRFs model, the model is divided into upper and lower layers and a gymnastic image filter that based on bag-of-visual-words semantic model. Identifying the error action image by the semantic, not only in line with the cognitive process of machine vision, and can effectively compensate the existing algorithm correcting the high false positives rate. Experiments show that by combining the two algorithms we can detect errors gymnastics image effectively, recognition rate compared with other algorithms improved, and the false detection rate reduced.

Keywords*: behavior recognition, Conditional Random Field (CRF), bag-of-visual-words, semantic model*

1. Introduction

With the popularization and development of artistic gymnastics, it is not only an important content in college physical education, but also an important way and means of the current mass sports. Artistic gymnastics was included in university, middle school and primary school curriculum, and a formal teaching content to teach, so that more people join the exercise of artistic gymnastics. Different courses may have different influence and role on female students' parts such as aerobics, physical training are aerobic exercise. Therefore aerobics, physical training can cause weight loss. Modern Aerobics basic pay more attention on the foot pace and movement, the movement of the upper limbs is just a simple action, and are mostly involved in the movement of large muscle groups, so complex class may reduce weight and arm circumference. Physical training to students who has poor pose and habits that lead to poor physical development also a great help, such as: O leg, X leg, spinal curvature, hunchback, *etc*., have a good repair.

Behavior recognition in human-interaction, visual tracking, robotics, computer vision is a very active and promising research [1-2]. Quickly and efficiently detect and identify the gymnastic action has become an important research for athletes training and game rule in recent years. Existing gymnastics action detection algorithm has good detection and recognition results but cannot effectively identify a variety of consecutive gymnastic actions and many gymnastics has high false rate.

CRFs (Conditional Random Field, CRF) model [3-4] achieved global normalized distribution in the entire state space and fusion context feature into the model, but it cannot express the substructure of the sequences. To solve this problem, [5] proposed Hidden CRF model, HCRF, which achieved the Spatial Correlation modeling, but the

ISSN: 2005-4254 IJSIP

model it is an offline model. To solve the problem of real-time of HCRF model [6] proposed Latent-Dynamic CRF model, LDCRF, the model training the dynamic characteristics and can mark undivided sequence pose directly but LDCRF model has deviant when the behavior converting and it cannot show the visualize action change.

In addition, Bag-of-Visual-Words algorithm is successfully used in image retrieval and scene classification fields [7-10] add information on this basis and propose feature descriptor HUE-SIFT, which fixed vocabulary size distribution, and finally use the classification algorithm SVM (Support Vector Machine) to classify images. But the algorithm lacks optimization itself, and needs much computing time. [11] proposed to build SURF visual vocabulary to extract local feature descriptors of the skin-color region so that we can reduce the computation time of SIFT algorithm [12], although the computational efficiency of SURF has improved greatly than SIFT algorithm, but the clustering process using a simple K-Mean clustering algorithm, resulting the lack of semantics.

Both methods for target identifying all exists flaw. So on this paper we improve the above two methods and combine the advantages of both models to build a hierarchical model for behavior recognition(as shown in the Figure 1), first we create a hierarchical semantic mark CRFs model, the model is divided into upper and lower layers, training respectively, identifying jointly, correction complementary, then the model is applied to human behavior recognition and achieved better recognition results. As image filtering algorithms has high rate of false positives, so that we establish a gymnastic image filter that based on bag-of-visual-words semantic model. The model based on the traditional low-level visual features, using the semantic analysis and understanding of the relevant technology to extract high-level semantic features contained in the image. Identifying the error action image by the semantic, not only in line with the cognitive process of machine vision, and can effectively compensate the existing algorithm correcting the high false positives rate. Experiments show that by combining the two algorithms we can detect errors gymnastics image effectively, recognition rate compared with other algorithms improved, and the false detection rate reduced.

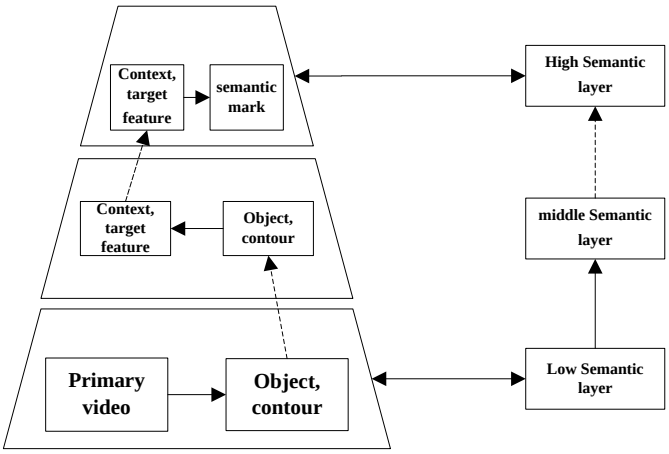

Figure 1. Hierarchical Model for Behavior Recognition

2. Stratification Semantic Mark CRFs Model

For the problems of the LDCRF model, we propose a hierarchical semantic mark CRFs (SMCRF) model, which improves LDCRF model; Figure 2 shows the SMCRF training and recognition processes. The database is divided into two parts: one for the use of the full video sequence with primary mark, the other acts as a single sequence with semantic mark. This paper presents the concept of hierarchical marks; the actions are subdivided into several representatives to establish the action structure. We use the model to handle

the video from the first layer (*i.e.*, top layer) to the second layer (lower layer) for behavioral recognition and the result is a set of middle-level semantic mark. Then we use the directivity and integrity of the semantic mark to identify the continuity of the behavior in this video and correct the recognition results, thereby enhancing the ability to identify human behavior.

Top layer of SMCRF model is used for class behavior recognizes and the lower layer is used for semantic mark the video sequences and detect the integrity and directivity. The top class recognition to extract contour features of the sequence databases to train structure parameter model then it can be used for the class test, the test results will be output as a lower input; segmenting single complete behavior video sequences from the database, training the lower layer structural parameters, using the upper output as the test inputs and get the corresponding video sequences, mark it, analyze the results and correction it. When in the testing stage the top layer is decisive to the output of the lower layer it determine the overall class; the lower will complete the top and make the output more accurate. Two together will make the behavior recognition more accurate, intuitive and consistent.

While the model is hierarchy, but one, the structure diagram shown in Figure 2, where x_i shows the i-th observation (corresponding to the i-th frame of the video sequence), h_i is a hidden state to x_i, different from with LDCRF we use y_i represent the semantic mark of x_i.

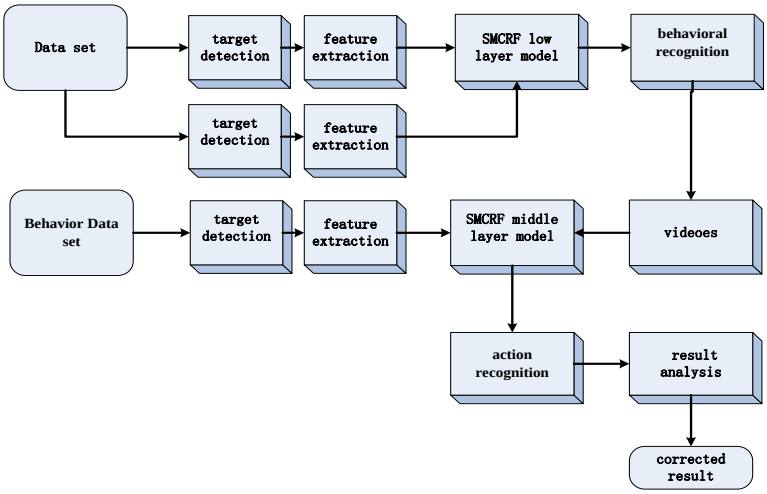

Figure 2. Lower and Middle Layer Structure of the Behavior Recognition

2.1 Hierarchical Semantic Mark Conditional Random based Behavior Model

This article we analyze the motion video sequences to identify the type of behavior and automatically verify the integrity and directivity online. The integrity of behavior refers to the various actions that compose the behavior in time-series order, which called a behavior class. Such as gymnastics accord with a certain speed to complete various gymnastics poses. Directivity refers to the actions that compose the behavior has specific context and cannot change the location we use this ideology modeling the SMCRF.

From observation sequence $X = (x_1, x_2, \ldots x_m)$ and the mark sequence $Y = (y_1, y_2, \ldots y_m)$ we get a map set. y_i is a class mark of i-th frame in each video sequence and the member of Y. Each x_i use feature vector$\varphi(x_i) \in R^d$ to express. Each frame of the sequence assumes the sub-structure variable vector $H = (h_1, h_2, \ldots h_m)$. These variables cannot be observed from the training sample, therefore built a hidden variables set in the model. Based on the assumptions above, we define a hidden condition random field model:

$$\mathrm{P(y|x,\theta)} = \sum_{h_i} P(y|h_i, x, \theta)\, P(h_i|x, \theta) \quad (1)$$

θ represents the model parameter. For effective training and learning, we need to restrict the model so that it will not contain a hidden state that associated with each class mark. For y_j each h_i is included in the set of all possible hidden state setH_{y_j}.For any sequences that $h_j \notin \mathrm{H}_{y_j}$, $\mathrm{P(y|x,\theta)} = 0$, model is as follows:

$$\mathrm{P(y|x,\theta)} = \sum_{h_j \in \mathrm{H}_{y_j}} P(h_i|x, \theta) \quad (2)$$

Among them, $\mathrm{P(y|x,\theta)}$ use the normal CRF:

$$\mathrm{P(y|x,\theta)} = \exp\left(\sum_k \theta_k \mathrm{F}_k(\mathrm{h,x})\right) / \mathrm{Z(x,\theta)} \quad (3)$$

$$\mathrm{Z(x,\theta)} = \sum_h \exp\left(\sum_k \theta_k \mathrm{F}_k(\mathrm{h,x})\right) \quad (4)$$

The expression of $\mathrm{F_k(h,x)}$ is as follows:

$$\mathrm{F}_k(\mathrm{h,x}) = \sum_{j=1}^{m} f_k\,(h_{j-1}, h_j, x, j) \quad (5)$$

Where $\mathrm{Z(x,\theta)}$is the partition function, k is the number of features, characteristic function $f_k(h_{j-1}, h_j, x, j)$ is a state function $S_k(h_j, x, j)$ or switching function$t_k(h_{j-1}, h_j, x, j)$. State function $S_k$is determined by each latent variable of the model, switching function t_k is decided by the hidden variables pair.

2.2 Top Parameter Training and Behavior Recognition

According to Figure 2 train the SMCRF model. Training set is divided into two types: one is the primary mark, one is semantic mark. The training set of primary mark is used for training the parameters of the first layer and overall classification of the video sequence; semantic mark parameters is used for training is the second layer, testing and calibration the video sequence. Primary mark training set consists of n marker sequences(x_i, y_i)(i = 1,2, ... n). θ^*[6] get from the following objective function:

$$\mathrm{L(\theta)} = \sum_{i=1}^{n} lnP(y_i|x_i, \theta) - \frac{1}{2\sigma^2}\|\theta\|^2 \quad (6)$$

Where: $\sum_{i=1}^{n} lnP(y_i|x_i, \theta)$is a condition log-likelihood values of the training sample, $\frac{1}{2\sigma^2}\|\theta\|^2$ is the Gaussian logarithmic of priori variance σ^2. We use Gradient descent algorithm find the optimal parameter values, so that$\theta^* = arg\max_\theta L(\theta)$. According equation (2)-(3), and the parameters θ_k that associated with the state function S_k is corresponding to a single training sequence(x_i, y_i), $lnP(y_i|x_i, \theta)$:

$$lnP(y_i|x_i, \theta) = \sum P\left(h_j = a\middle|y, x, \theta\right) S_k(j, a, x) - \sum P\left(h_j = a, y'\middle|x, \theta\right) S_k(j, a, x) \quad (7)$$

Wherein the edge gradient probability $P\left(h_j = a\middle|y, x, \theta\right)$ can be calculated by the Belief Propagation [6]:

$$P\left(h_j = a\middle|y, x, \theta\right) = \frac{\sum_{h:h_j = a \wedge h_j} p(h|x,\theta)}{\sum_{h:h_j} p(h|x,\theta)} \quad (8)$$

When SMCRF model lower recognition result has larger deviation with the recognition results of the upper, *i.e.*, the upper classification results inconsistent with its corresponding structure, so we need to compare it to other classes mark, through experiments when the probability by the formula (7) was greater, the corresponding marker and detection sequence has the higher the degree of structure consistent, given some test sequence, this sequence mark obtained by the following formula:

$$\mathrm{y} = \arg\max_y p(y|x, \theta^*) \quad (9)$$

Where θ^* obtained from the training samples. Each class mark is related to series of

disjoint hidden state, so formula (9) can be written as:

$$y = \arg\max_y \sum_{h:h_j} p(h|x, \theta^*) \quad (10)$$

To estimate the j-th frame marky_j, calculated edge probability $P(h_j = a|y, x, \theta^*)$ for all hidden state a ∈ H, the mark obtained by the above method enables the error is minimized for each frame.

3. Bag of Visual Words Semantic Model

We proposed a bag of visual word semantic model shown in Figure 3. The model includes:

1) semantic dictionary constructed by Random Forest algorithm [7] which extract visual primitives to classify the underlying visual vocabulary and vocabulary again intermediate classify, integrate context and spatial feature [13-15], then we will build a visual semantic dictionary with high degree distinction;

2) the classifier use[7] The x^2-kernel SVM classification algorithm.

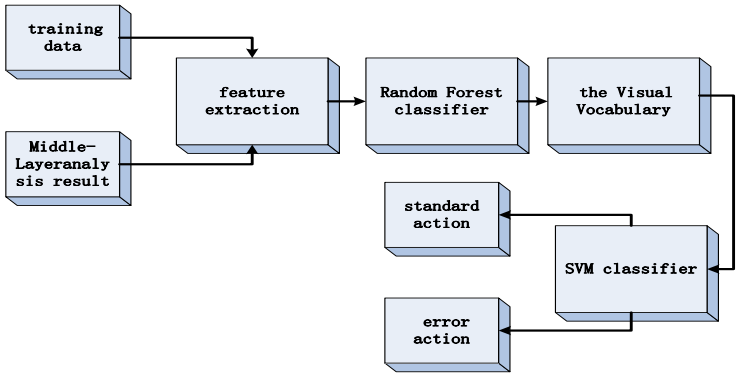

Figure 3. Bag of Visual Word Semantic Model

3.1. The High-Level Semantic Dictionary Construction Algorithm

The high-level semantic dictionary construction algorithm is mainly constructed through visual classification and classification of visual words two phases. Visual primitive classification algorithm consists of random forest algorithms and interference cancellation algorithm to complete; visual word classification is completed by the contextual visual vocabulary allocation algorithm and spatially correlated visual vocabulary allocation algorithm. Finally we will complete the construction of semantic visual dictionary. The constructing of the semantic dictionary algorithm is shown in Figure 4.

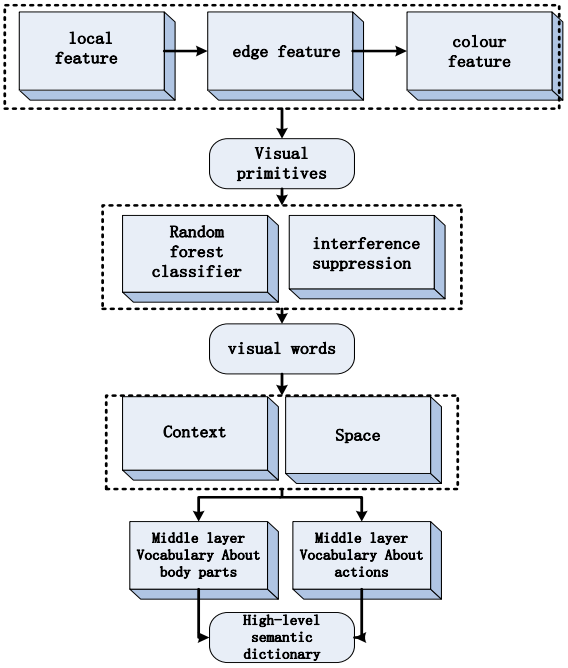

Figure 4. The Constructing of the Semantic Dictionary Algorithm

3.2. Weight Distribution of Spatially Correlated Visual Vocabulary

The most obvious feature of gymnastics is a large area of the skin, if we weighted distribution of the visual vocabulary which is related with the gymnastics actions space during the process of the visual vocabulary distribution, it can effectively increase the identification of the object areas and the visual discrimination. In this paper we weighted distribution the visual vocabulary that correlated with the gymnastics actions space [13]. Algorithm is described as follows:

1) Visual vocabulary set that extract from the interest points set $w_l(1 < l < k, k$ is the number of the visual vocabulary in the visual dictionary);

2) The visual vocabulary w_l weighted frequency $T_{fw_l\beta_i}$ under Gaussian β_i;

$$T_{fw_l\beta_i} = \sum_{m=1}^{n_i} P(\beta_i / Z_m) \tag{11}$$

3) Calculate average weights of the visual vocabulary w_l;

$$T_{fw_l} = \sum_{i=1}^{n_{w_l}} \frac{T_{fw_l\beta_i}}{n_{w_l}} \tag{12}$$

4) Calculate the weight of w_l;

$$I_{fw_l} = ln \frac{n}{n_{w_l}} \tag{13}$$

5) Calculate the final spatial weights of visual vocabulary w_l.

$$S_{w_{w_l}} = T_{fw_l} \times I_{fw_l} \tag{14}$$

3.3. The Body Parts Feature Space Topology Algorithm

Traditional bag of visual vocabulary algorithm encoding the visual primitives of the local prominent visual fragment to get the visual vocabulary, but they ignore the topological structure of between the images; we use the space topology feature of the body part to express the high distinction degrees characteristics of the gymnastics. The study found that the body part itself has a certain characteristic of spatial topology; using

the characteristic features of the body parts we can determine whether the gymnastics actions are standard. For the gymnastic actions that body inclination is not greater than 90°, human torso and limbs compose with geometric shapes.

In this paper, we use simplified 2D string [14] to build the space topology structure between the semantic visual vocabulary, assuming that V is the feature set of human body parts, each object corresponds to a high-level visual vocabulary, this paper take the "<" to show the space around or up and down relationship. We find out that the body parts form a string of "B <A <C". The images which meet such relationships will be recognized as wrong action of gymnastics.

3.4. High-level Semantic Tree Algorithm

For gymnastics not only contain skin color information and more importantly it contains some specific actions information, such as falls, collisions and the action is not in place. Fall mainly related to the torso and limbs; collision mainly related to the limbs; and the action is not in place mainly related to the human torso and limbs. The semantic of the above three behaviors can compose semantic features in gymnastics. The high-level semantic tree construct process of gymnastics is shown in Figure 5.

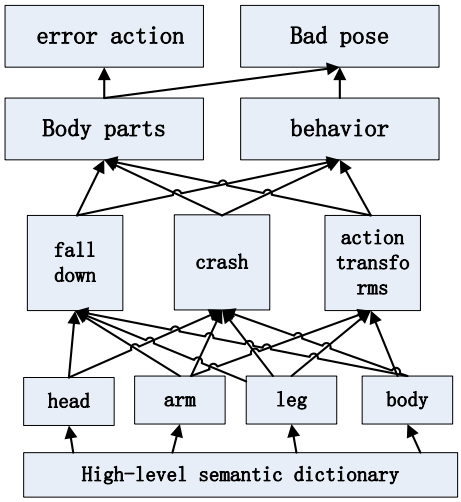

Figure 5. The High-Level Semantic Tree Constructs

In Figure 5, the semantic model comprising: body parts model and the human body behavior model. Due to some words of gymnastics semantic vocabulary has high appearing probability, therefore, in this paper on the basis of [15] they weighted distribution the semantic vocabulary we proposed to removal of the scene semantic model and only weight distribution the human body parts visual vocabulary and human behavior related visual vocabulary. Any image C_i in data images set V, if the visual vocabulary probability of the body parts is P_0 human behavior related visual vocabulary probability is P_A, by the formula $P_{C_i} = \alpha P_0 + \beta P_A$, $\alpha + \beta = 1$,we can calculate the error rate of C_i. The image shows the wrong action when the error rate is greater than the threshold t.

4. Experimental Results and Analysis

We use HCRF tool bag, features extracted by VC ++and Matlab completed the training and testing. Video data come from the Weizmann Human Action Dataset [15] and network video. Our algorithm has semantic mark, training is divided into three parts, the upper layer complete the primary semantic mark model training, middle layer complete the middle semantic mark training. The high-level semantic model is classify by the contextual and spatial correlation characteristics and further classify semantic information

from the middle layer, the semantics vocabulary was further grouped into 2000 body parts and 3000 body movements-related high-level semantic vocabulary. Finally, we encode the high-level semantic vocabulary to form the semantic dictionary. Since the error actions included a large area of skin and body parts, such local feature points have significant context and spatial correlation properties.

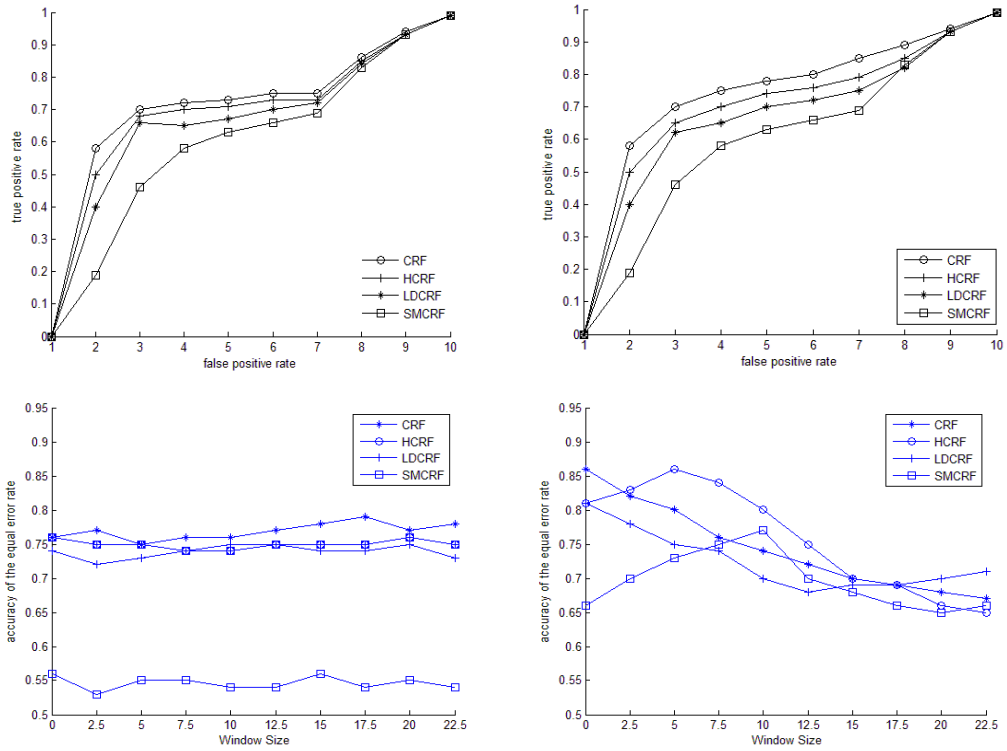

Figure 6. Statistics of Four Models EER (Equal Error Rate)

4.1. Comparisons of Experimental Data

Experiments using K-fold cross to verify the model, *i.e.*, each time we use K-1 sequence database as the training samples and the other as the test samples. Experiment is under different size of windows: w =0 represents the current frame, w = 1 is meant to a total of three which include the current frame and the back and forth frames image data, and so on. In the experiment, we compare the CRF, HCRF, LDCRF modeling, validation the identification capability of SMCRF. Figure 6 shows the ROC curve of the gymnastics and running with four models on the window of 0. The figure shows that the identification capability of SMCRF model is superior to CRF, HCRF and LDCRF model. Figure 6 statistics of four models EER (Equal Error Rate) recognition accuracy under different window size, we can see that the effect is not to turn better with the increase of the window, such as LDCRF, HCRF model will change before or after the window at 10, and SMCRF model identify better than other models when the window is small, as the window becomes larger, this advantage will slowly disappear. This is because as the sequence span becomes large, the number of class will become bigger, so that the calculation speed of the entire model decrease and the output is complex, through experiment we find that the SMCRF model has superior results when the window between 3-5.

Table 1. Accuracy of Three Types of Classification Algorithm %

method	fall down	crash	action transforms
Literature[6]	86.1	73.0	66.7
Literature[4]	82.3	86.2	64.8
Our method	92.8	91.0	86.5

Table 1 the results show that: the algorithm added contextual and spatial correlation visual vocabulary feature has better distinguish than the traditional visual allocation algorithm; the correct rate is higher than the other three image filtering algorithms. The time complexity of bag of visual words algorithm depends on the time complexity of feature extraction and distribution of visual vocabulary algorithm, we use the Random Forest algorithm of [7] to assign visual vocabulary and the algorithm shorten the computation time. Experimental results show that the total time of feature extraction and vocabulary allocation two-stage and the average computation time of the proposed method are lower than the other three algorithms. Four types of feature extraction algorithms and vocabulary allocation average computation time is shown in Figure 6.

Comprehensive Figure 6 we can find that the SMCRF, LDCRF, HCRF that contain hidden variables are significantly better than the model without hidden state CRF. When the window is small, the recognition result of SMCRF is preferably, because SMCRF add of the authentication mechanism on the basis of LDCRF, so that it reduces the recognition error rate effectively.

5. Conclusions

In this paper, we proposed a hierarchical semantic model based on conditional random fields, obtain the dynamic characteristics and hidden dynamic characteristics of behavior, so that it can handle video segmentation and labeling work; introduces a hidden state variable based on frames, and modeling the sub-structural behavior feature so that obtain the dynamic characteristics of behavior marks. By contrast CRF, HCRF, LDCRF model behavior recognition sequence segments capability, we verified the SMCRF that we proposed has better modeling capability, and online recognition capability. Among them, the detection based on bag of visual words is a promising approach because of the error action was filtered through a high-level semantic understanding of image expression not only can filter apparent error action, but also in line with machine vision and cognitive theory. In this paper, high-level semantic features with a high degree of distinction and the classify detection speed and accuracy with the existing algorithms have greatly improved, but there are many challenging issues remain unresolved, such as extract of the same behavior efficient and accurate from different angles, improve the speed of model training, *etc*. These issues will be the focus of the future work.

References

[1] Y. Hejin and W. Cuiru, "Human Action Recognition Using Markov Random Walk Based Semi-Supervised Learing", Journal of Computer-Aided Design & Computer Graphics, vol. 23, no. 10, **(2011)**, pp. 1749-1757.
[2] G. Jun-xia, D. Xiao-qing and W. Sheng-jin, "A Survey of Activity Analysis Algorithms", Journal of Image and Graphics, vol. 14, no. 3, **(2009)**, pp. 377-387.
[3] L. Y. Zhao, X. Wang and G. Sukthankar, "Motif discovery and feature selection for CRF-based activity recognition", 20th International Conference on Pattern Recognition, Piscataway, **(2010)**.
[4] H. Lei, L. Junfeng and J. Yunde, "Human Interaction Recognition Using Spatio-Temporal Words", Chinese Journal of Computers, vol. 33, no. 4, **(2010)**, pp. 776-784.
[5] S. B. Wang and A. Quattoni, "Hidden conditional random fields for gesture recognition", 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, **(2006)**.
[6] L. P. Morency, A. Qunattoni and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition", 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, **(2007)**.

[7] J. R. R. Uijlings, A. W. H. Smeulders and R. J. H. Scha, "Real-time visual concept classification", IEEE Transactions on Multimedia, vol. 12, no. 7, **(2010)**, pp. 665-681.
[8] W. Lei, S. C. H. Hoi and Y. Nenghai, "Semantics-preserving bag-of-words models and applications", IEEE Transactions on Image Processing, vol. 19, no. 7, **(2010)**, pp. 1908-1920.
[9] A. P. B. Lopes, de Avila S E F, Peixoto A N A, et al. A bag-of-features approach based on HUE-SIFT descriptor for nude detection, EUSIPCO 2009: Proceedings of the 17th European Signal Processing Conference. (2009) Glasgow, Scotland.
[10] Lopes A P B, S. E. F. de Avila and A. N. A. Peixoto, "Nude detection in video using bag-of-visual-features", SIBGRAPI'09:Proceedings of the 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing, Washington, DC, **(2009)**.
[11] L. Yizhi and X. Hongtao, "Constructing SURF visual-words for pornographic images detection", ICCIT'09:12th International Conference on Computers and Information Technology, Washington, DC, **(2009)**.
[12] L. Juan and O. Gwon, "A comparison of SIFT, PCA-SIFT and SURF", International Journal of Image Processing, vol. 3, no. 4, **(2009)**, pp. 143-152.
[13] L. Teng, M. Tao and I. Kweon, "Contextual bag-of-words for visual categorization", IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 4, **(2010)**, pp. 381-392.
[14] W. Mei, W. Yanling and L. Guangda, "Object recognition via adaptive multi-level feature integration", APWEB 2010: 12th International Asia-Pacific Web Conference, Washington, DC, **(2010)**.
[15] C. Mianshu, F. Ping and L. Yong, "Condensed semantic tree model for image category representation", 2010 ICCAE: Proceedings of the 2nd International Conference on Computer and Automation Engineering, Chongqing, **(2010)**.

