

# Action recognition using a hierarchy of feature groups

Zhou Tongchi Cheng Xu Li Nijun Xu Qinjun Zhou Lin Wu Zhenyang

(School of Information Science and Engineering, Southeast University, Nanjing 210096, China)

**Abstract:** To improve the recognition performance of video human actions, an approach that models the video actions in a hierarchical way is proposed. This hierarchical model summarizes the action contents with different spatio-temporal domains according to the properties of human body movement. First, the temporal gradient combined with the constraint of coherent motion pattern is utilized to extract stable and dense motion features that are viewed as point features, then the mean-shift clustering algorithm with the adaptive scale kernel is used to label these features. After pooling the features with the same label to generate part-based representation, the visual word responses within one large scale volume are collected as video object representation. On the benchmark KTH (Kungliga Tekniska Högskolan) and UCF (University of Central Florida)-sports action datasets, the experimental results show that the proposed method enhances the representative and discriminative power of action features, and improves recognition rates. Compared with other related literature, the proposed method obtains superior performance.

**Key words:** action recognition; coherent motion pattern; feature groups; part-based representation

**doi:** 10.3969/j.issn.1003-7985.2015.03.005

Human action recognition (HAR) is one of the hot topics in the fields of computer vision and pattern recognition due to its widespread applications in video surveillance, human computer interaction and video retrieval. However, its research is influenced by significant camera motion, background clutter, and changes in object appearance, scale, illumination conditions and viewpoint. Overall, HAR has become a difficult but also an important task.

Local features together with bag-of-visual words<sup>[1-3]</sup> (BoVW) have gained good recognition performance. Kovashka et al.<sup>[1]</sup> employed the Euclidean metric to construct variable-sized configurations of local features and learned compound features, and each action video is modelled by the learned compound features in a hierarchical way. Also, Yuan et al.<sup>[4]</sup> used the same metric to

measure the distance between features, and counted the co-occurrence frequency of pair features within some spatial-temporal extents. Considering the activity data containing information at various temporal resolutions, Song et al.<sup>[5]</sup> presented a hierarchical sequence summarization and learned multiple layers of discriminative feature representations. In fact, the methods in Refs. [1, 4] with the popular spatio-temporal interest points (STIPs), like cuboids, and 3D Harris etc. are easily influenced by the camera motion and background clutter, so the learned context<sup>[1-5]</sup> lacks the representativeness. To extract stable features for action recognition, the motion compensation technique<sup>[6]</sup> is introduced to suppress the camera motion. Chakraborty et al.<sup>[7]</sup> selected STIPs by surrounding suppression combined with local and temporal constraints. Moreover, to reduce the quantization error and preserve the nonlinear manifold structure, Refs. [8–9] adopted structured sparse coding to encode the local features for recognition tasks.

Inspired by the ideals of Refs. [1, 5–8], we learn a spatial-temporal context with an ascending order of abstraction in a hierarchical way. We first compensate camera motion, and then utilize temporal gradients to extract stable motion features. To learn local context and model body parts, we utilize the clustering algorithm instead of the ranked metric. After encoding the underlying features with locality- and group-sensitive sparse representation (LGSR)<sup>[9]</sup> and learning part-based representation, the large scale context for the constructed volumetric region is sequentially modelled. From experiments, our representation enhances the discriminative power of action features and achieves excellent recognition performance on the benchmark KTH (Kungliga Tekniska Högskolan) and UCF (University of Central Florida)-sports action datasets.

## 1 Proposed Method

The hierarchical feature representation model for action recognition proposed in this paper is semantic structures from motion including region, part and object, as shown in Fig. 1(c). The initial layer is the low-level features extracted from salient 3D motion regions. The second layer is a pool group features labelled by the mean-shift clustering algorithm. The top layer is explored to construct body movement representations. Using our method to represent the irregular 3D regions is more flexible than those with fixed grids<sup>[10]</sup> or nearest rank<sup>[11,4]</sup>, as shown in

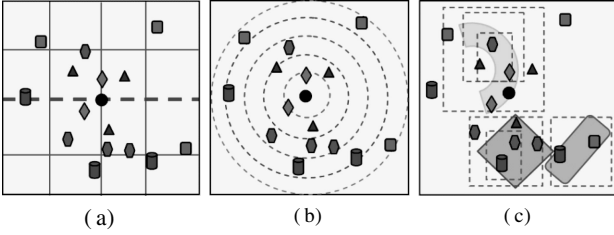
**Received** 2015-01-04.

**Biographies:** Zhou Tongchi (1979—), male, graduate; Wu Zhenyang (corresponding author), male, doctor, professor, zhenyang@seu.edu.cn.

**Foundation item:** The National Natural Science Foundation of China (No. 60971098, 61201345).

**Citation:** Zhou Tongchi, Cheng Xu, Li Nijun, et al. Action recognition using a hierarchy of feature groups[J]. Journal of Southeast University (English Edition), 2015, 31(3): 327 – 332. [doi: 10.3969/j.issn.1003-7985.2015.03.005]

Figs. 1(a) and (b), respectively.



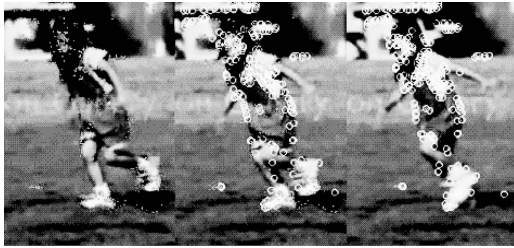
**Fig. 1** The learned spatial-temporal relationships. (a) Multi-level fixed grids<sup>[10]</sup>; (b) Nearest rank<sup>[1,4]</sup>; (c) Our method

### 1.1 Extracting and encoding features

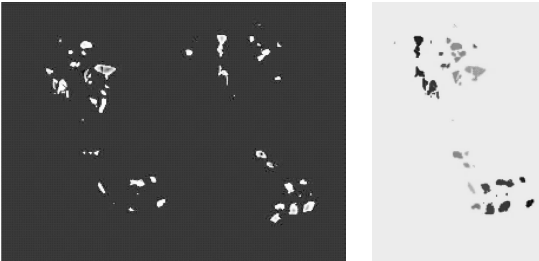
Each video is first segmented to narrow clips. According to the motion compensation<sup>[6]</sup>, the 2D polynomial affine motion models are considered for estimating the dominant motion, and the adjacent frames from a narrow clip are aligned. Fig. 2(a) shows an example with the high score matching corner points in the region of interest (ROI), where the corresponding dots are in the first and second columns, and the circles in the second and third columns. After the processes of motion compensation, temporal gradient and the adaptive threshold, the saliency map shown in Fig. 2(b) is obtained by

$$D_i = \begin{cases} \text{abs}(I_i - I'_{i-1}) & \text{if } D_i(x, y) \geq 0.5v \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where  $D_i(x, y)$  means the pixel value;  $I'_{i-1}$  is the compensated ROI image corresponding to  $I_{i-1}$ ; and  $v$  is the saliency maximum of the absolute difference image  $D_i$ .



(a)



(b)

(c)

**Fig. 2** Processes of saliency motion extraction. (a) Matching points; (b) Difference images; (c) Saliency motion cumulative image

Unfortunately, some noise still persists, and the process mentioned above may lead to leak detection. To handle these problems, the constraints of some context clues including coherent motion patterns, changes of dis-

placement and phase, color value invariance in short duration are adopted to select saliency motion sub-regions. According to the central coordinate set  $\{r'_1, r'_2, \dots, r'_n\}$  obtained by 8-connected regions of the difference cumulative image, we sample sub-volume within the difference volume constructed by difference images. For one sub-volume, the spatial window for KTH and UCF-sports is empirically set to be  $10 \times 10$  and  $30 \times 30$ , respectively, and temporal scale  $L$  is the number of difference images. The process to extract motion features is followed by

$$w_i = \begin{cases} 1 & \text{if } \exists P_i(x, y) > 0, i = 1, 2, \dots, L \\ 0 & \text{otherwise} \end{cases}$$

$$M(w_1, \dots, w_i, \dots, w_L) = \begin{cases} 1 & \text{if } \text{sum}(M) \geq 2, \exists i, i+1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where  $P_i(x, y)$  is the pixel value at coordinate  $(x, y)$  of the  $i$ -th patch of a sub-volume;  $w$  and  $M$  are defined as the Boolean-valued function;  $w$  is used to distinguish whether the motion exists in some patch and  $M$  is used to distinguish whether the spatial-temporal feature is valid. After pruning, the new center coordinate set  $\{r_1, r_2, \dots, r_m\}$  is viewed as STIPs, and the new difference cumulative image shown in Fig. 2(c) is defined as

$$T_v = \sum_{p=1}^L D'_p \quad (3)$$

where  $D'_p$  means the difference image after pruning  $D_p$ .

After describing volumes, we transform the HoG/HoF descriptors (HoG: 4 bins; HoF: 5 bins) into structured sparse representations by LGSR. This encoding method takes advantages from both group sparsity and data locality structure in determining the discriminative representation for classification<sup>[9]</sup>.  $C_g$  can be solved by the following optimization problem:

$$\min_{C_g} \left\| H_g - DC_g \right\|_2^2 + \lambda_1 \sum_{j=1}^n \left\| C_{g,j} \right\|_2 + \lambda_2 \left\| v \Theta C_g \right\|_2^2 \quad (4)$$

where  $D = [D_1, D_2, \dots, D_n] \in \mathbf{R}^{D \times d}$  is the codebook;  $\lambda_1$  and  $\lambda_2$  are the weights for the group sparsity and locality constraints, respectively; and the vector  $v \in \mathbf{R}^{d \times 1}$  is the distance measurement between  $H_{g,j}$  and each visual word.

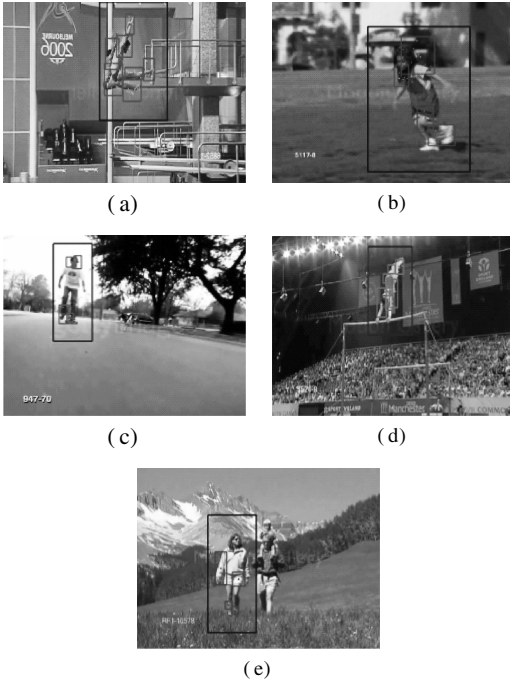
### 1.2 Group feature generation by clustering

In this section, we use the mean-shift clustering algorithm to construct group features, and then adopt a max-pooling operator to generate part-based representation. For the mean-shift clustering, 3D template with the adaptive scale is used instead of the fixed bandwidth kernel that is unsuitable to model the irregular movement part. The temporal scale of 3D kernel is  $L$  frames. The 3D kernel's spatial scale  $(r_x, r_y)$  controlling the ranges of motion sub-region centers is a parameterized function, which can adaptively change by zooming in or out of the

body scope ( $O_{-x}$ ,  $O_{-y}$ ) as follows by the given annotation. The scale ( $r_x$ ,  $r_y$ ) is defined as

$$r_x = \frac{r_{\text{ref}-x} O_{-x}}{R_{-x}}, \quad r_y = \frac{r_{\text{ref}-y} O_{-y}}{R_{-y}} \quad (5)$$

where  $R_{-x}$  and  $R_{-y}$  are set to be (80, 50), and they are the reference sizes of body scope;  $r_{\text{ref}-x}$  and  $r_{\text{ref}-y}$ , the spatial sizes of template, are set to be (20, 20) in our experiments. Moreover, if the features deviate from the clustering center, the extracted color information (gray image: illumination) is used to re-label them. After the above two stages, some action features with the same label can give an enough coverage of the body part, as shown in Fig.3, where the defined ROIs are represented by large blue boxes. The body movement parts are described by small red boxes, and the centers of 8-connected motion regions are denoted by green points in red boxes.



**Fig. 3** Clustering results of motion features for some action video frames. (a) Dive; (b) Kick; (c) Skate; (d) Swing 2; (e) Walk

After clustering, the coefficient set  $C_g \in \mathbf{R}^{d \times k}$  corresponding to the descriptor set  $H_g$  is represented as

$$C_g = \{C_{g,1}, C_{g,2}, \dots, C_{g,k}\} \quad k = 1, 2, \dots, m \quad (6)$$

The max-pooling<sup>[10]</sup> operator for  $C_g$  is defined by

$$S(i) = \max(\text{abs}\{C_{g,1}(i), \dots, C_{g,k}(i)\}) \quad i = 1, 2, \dots, d \quad (7)$$

where  $S(i)$  represents the maximum absolute response of the  $i$ -th atom, and  $S$  is the descriptor for certain body part.

### 1.3 Object-level context

Due to the limited scale of body parts, it is not enough to capture large scale co-occurrence relationships. We use ROIs of narrow clips to construct volumes which can

adaptively adjust the spatial scales following the changing human body scope, and then accumulate each element of all part descriptors as volume descriptors. The produced vector is computed as

$$V(i) = \sum_{g=1}^n S_g(i) \quad (8)$$

where  $V(i)$  is the weight accumulation of the  $i$ -th atom response, and  $S_g$  denotes the  $g$ -th body part descriptor.

## 2 Action Representation and Recognition

After describing feature groups and object context, each video is represented by the descriptors of linear quantization corresponding to different levels, and the lengths are all  $N_{\text{atoms}} N_{\text{bin}}$ , where  $N_{\text{atoms}}$  denotes the dictionary size, and  $N_{\text{bin}}$  represents the quantization bins.

Recognition is performed by the nearest neighbour classifier (NNC) and support vector machine (SVM). The NNC is a simple and effective classifier and the absolute distance is used to measure the similarity. For the SVM classifiers, we adopt  $\chi^2$  kernel<sup>[11]</sup> which is an extension form of  $\chi^2$  distance<sup>[11]</sup> and the Gaussian radial basis function<sup>[12]</sup> (G-RBF), respectively. The two kernels are commonly used for classification task. The Gaussian radial kernel and  $\chi^2$  kernel are, respectively, defined by

$$K_{\text{G-RBF}}(H_i, H_j) = e^{-r \|H_i - H_j\|_2^2} \quad (9)$$

$$K(H_i, H_j) = \frac{2H_i H_j}{H_i + H_j} \quad (10)$$

where  $H_i$  and  $H_j$  represent the histograms of video representations. In the cases of the G-RBF kernel, the  $r$  values are selected heuristically.

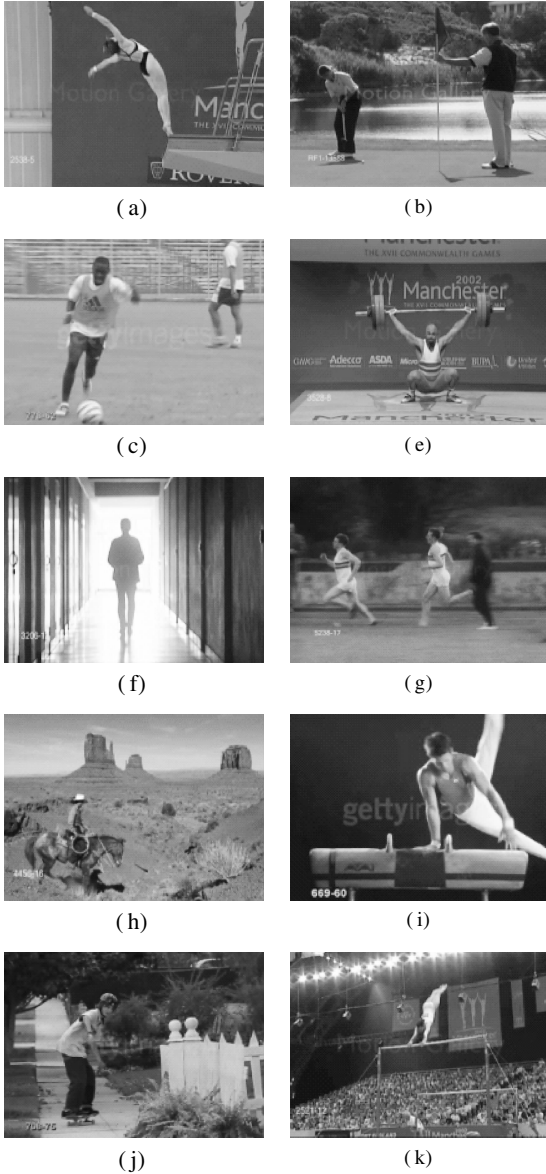
## 3 Experimental Results

### 3.1 Action dataset

We adopt KTH<sup>[1,7-8]</sup> and UCF-sports<sup>[1-2, 10,12-14]</sup> action datasets to validate our proposed method. The KTH dataset contains 599 videos of 25 actors performing six types of human actions, box, clap, wave, jog, run, and walk. Each action is repeated in four different scenarios: outdoors, outdoors with scale variation, outdoors with clothing variation and indoors. All sequences with low resolution are recorded. The UCF-sports dataset consists of 150 video clips acquired from sports broadcast networks. The videos have camera motion and jitter, highly cluttered and dynamic backgrounds, compression artifacts and variable illumination settings at variable spatial resolutions. Fig.4 shows the class of 10 action samples on the UCF-sports action dataset.

### 3.2 Experimental settings

The narrow clip length is set to be 3.  $\lambda_1$  and  $\lambda_2$  are set to be 0.3 and 0.1, respectively. Video frames in the KTH dataset have a simple background and slight camera



**Fig. 4** Sample action frames from video sequences of the UCF-sports dataset. (a) Dive; (b) Golf; (c) Kick; (e) Lift; (f) Walk; (g) Run; (h) Ride; (i) Swing 1; (j) Skate; (k) Swing 2

motion, so we do not need to align adjacent frames. The dictionary with 936 atoms is constructed by randomly selected 280 sets of group features. For the KTH dataset, we follow the leave-one-out cross-validation (LOOCV) evaluation scheme, and adopt the most simple NNC with  $k = 3$ . For UCF-sports dataset, the defined ROIs' scale is zoomed in 20% under the original center-frame ROI. The dictionary with 839 atoms is built by a randomly selected 105 group of descriptors. Our method is validated by the five-fold cross-validation<sup>[13-14]</sup> and the split evaluation scheme<sup>[2]</sup>. For the NNC, the neighbour parameter is set to be 5. With the SVM classifier, we adopt a one-against-rest training approach. For the G-RBF kernel, by cross-validation, the optimal values of two controlling parameters are set to be  $C = 380$  and  $r = 0.2$ . For the  $\chi^2$  kernel, the parameter  $C$  is set to be 380. The recognition accuracy is average result over 100 runs.

### 3.3 Evaluation on KTH dataset

Fig. 5 shows the recognition accuracy for the KTH dataset in the form of confusion matrix. From Fig. 5, the majority of the confusion between “jog” and “run” is expected due to similar nature between their local features. Tab. 1 shows performance comparison with other methods. Among the results<sup>[1, 7-8, 13]</sup> using local features to model actions, our method achieves 96.11% recognition accuracy.

Box	0.97	0.03	0	0	0	0
Clap	0.01	0.97	0.02	0	0	0
Wave	0	0.01	0.99	0	0	0
Jog	0	0.01	0	0.93	0.05	0.01
Run	0.01	0	0	0.03	0.94	0.02
Walk	0.02	0	0.01	0	0	0.97

**Fig. 5** Confusion matrix for KTH dataset

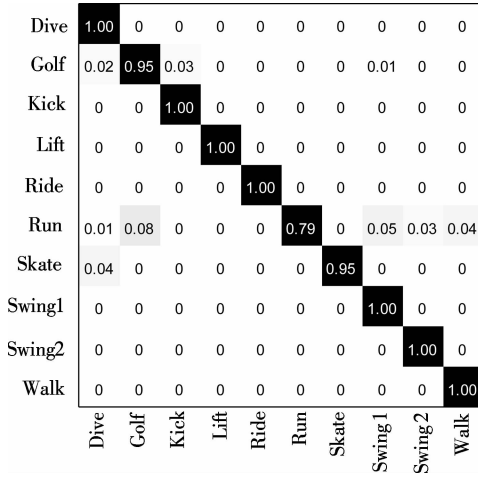
**Tab. 1** Performance comparison with other methods

Methods	Accuracy/%
Our method	96.11
Zhou et al. <sup>[8]</sup>	97.99
Chakraborty et al. <sup>[7]</sup>	96.35
Castrodad et al. <sup>[13]</sup>	97.60
Kovashka et al. <sup>[11]</sup>	94.50

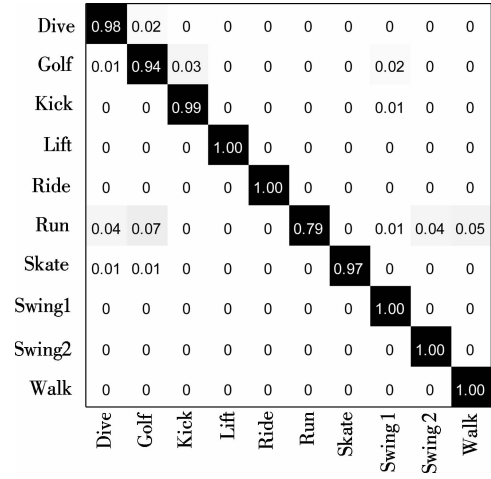
### 3.4 Evaluation on UCF-sports dataset

We first adopt the simple NNC to validate our proposed method. Under the five-fold cross-validation and split evaluation scheme, all class average recognition accuracies are shown in Figs. 6(a) and (b), respectively. For the SVM classifier, under the split scheme, the recognition results for all action videos are shown in Figs. 7(a) and (b) corresponding to the  $\chi^2$  kernel and G-RBF kernel, respectively. From Figs. 6 and 7, we can see that the majorities of recognition error are among “Golf”, “Skate” and “Run”.

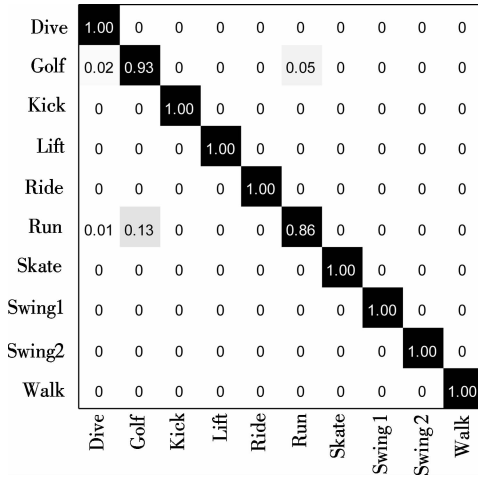
To evaluate performance at different levels with respect to histogram bins, we use the simple NNC with five-fold cross-validation manner. Fig. 8 shows the recognition accuracy plot varying with the histogram bins, where each point on the curves corresponds to an average result. At some bins, recognition accuracies with object representation are lower than that of the part representation, but recognition rates tend to be insensitive to the quantization bins. The recognition results with the part representation can reach 100% in some quantization bins.



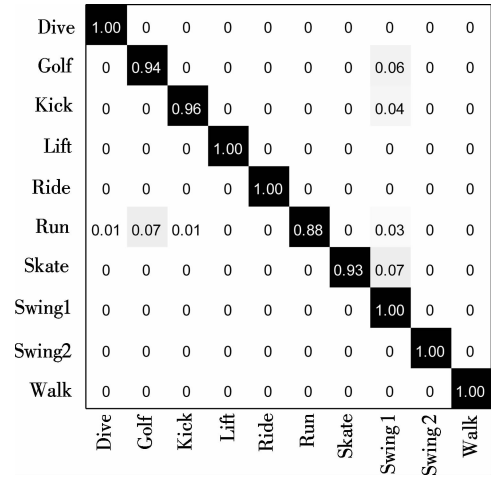
(a)



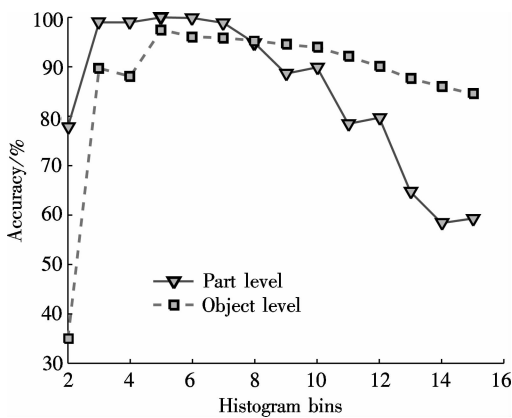
(b)

**Fig. 6** Confusion matrices with NNC for UCF-sports dataset. (a) Five-fold cross validation; (b) Splitting

(a)



(b)

**Fig. 7** Confusion matrices with SVM for UCF-sports dataset. (a)  $\chi^2$  kernel; (b) G-RBF kernel**Fig. 8** Recognition results at different levels

text and robust LGSR-based sparse representation. In addition, the recognition performance with the SVM classifier is better than that based on the NNC. Note that Sanin et al.<sup>[15]</sup> designed dense spatio-temporal covariance descriptors and adopted the LogitBoost classifier to recognize actions. Michalis et al.<sup>[14]</sup> utilized the dense trajectories to learn discriminative action parts in terms of an MRF score. Lan et al.<sup>[2]</sup> employed a figure-centric visual word representation for joint action localization and recognition.

**Tab. 2** Performance comparison with other methods

Methods	Classifier and settings	Accuracy/%
Our method	NNC: 5-fold/split	96.86/96.77
	SVM: $\chi^2$ /G-RBF	97.96/97.06
Sanin et al. <sup>[15]</sup>	LogitBoost (5-fold)	93.91
Castrodad et al. <sup>[13]</sup>	SVM (LOOCV)	97.3
Michalis et al. <sup>[14]</sup>	SVM (5-fold)	79.4
Lan et al. <sup>[2]</sup>	SVM (split)	73.1

#### 4 Conclusion

In this paper, we propose an action hierarchical model,

Tab. 2 lists performance comparison of our method with other methods on the UCF-sports dataset. Compared with the literature using local features to model actions, the recognition rates of our method using object representation are higher than those of other methods. The obtained better performance benefits from three aspects, having stable and dense motion features, a semantic con-

which can capture the discriminative statistics of co-occurring motion features at multiple levels. After extracting the stable and dense motion features by motion compensation techniques together with temporal gradient and coherent motion pattern constraints, we use the structured sparse representations of HoG/HoF descriptors as underlying features. Then the orderly hierarchical spatial-temporal context for different scale volumes is represented by aggregating group features generated by mean-shift clustering, and accumulating each element of visual word responses, respectively. On the KTH and UCF-sports action datasets, the experimental results show that our method obtains good performance.

## References

- [1] Kovashka A, Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition [C]//*Proc of the International Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA, 2010: 2046–2053.
- [2] Lan T, Wang Y, Mori G. Discriminative figure-centric models for joint action localization and recognition [C]//*Proc of the International Conference on Computer Vision*. Colorado, USA, 2011: 2003–2010.
- [3] Hu Q, Qin L, Huang Q, et al. Action recognition using spatial-temporal context [C]//*Proc of the 20th International Conference of Pattern Recognition*. Istanbul, Turkey, 2010: 1521–1524.
- [4] Yuan C, Hu W, Wang H, et al. Spatio-temporal proximity distribution kernels for action recognition [C]//*Proc of the International Conference of Acoustics, Speech and Signal Processing*. Dallas, TX, USA, 2010: 1126–1129.
- [5] Song Y, Morency L P, Davis R. Action recognition by hierarchical sequence summarization [C]//2013 *IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA, 2013: 3562–3569.
- [6] Jain M, Jegou H, Bouthemy P. Better exploiting motion for better action recognition [C]//*Proc of the International Conference of Computer Vision and Pattern Recognition*. Portland, OR, USA, 2013: 2555–2562.
- [7] Chakraborty B, Holte M B, Moeslund T B, et al. Selective spatio-temporal interest points [J]. *Computer Vision and Image Understanding*, 2012, **116**(3): 396–410.
- [8] Zhou T C, Chen X, Wu Z Y. Action recognition using hierarchically tree-structured dictionary encoding [J]. *Journal of Image and Graphics*, 2014, **19**(7): 1054–1061. (in Chinese)
- [9] Chao Y W, Yeh Y R, Chen Y W, et al. Locality-constrained group sparse representation for robust face recognition [C]//*Proc of the International Conference on Image Processing*. Brussels, Belgium, 2011: 761–764.
- [10] Xiao W H, Wang B, Liu Y, et al. Action recognition using feature position constrained linear coding [C]//*Proc of the International Conference on Multimedia and Expo*. San Jose, CA, USA, 2013: 1–6.
- [11] Vedaldi, A, Zisserman A. Efficient additive kernels via explicit feature maps [C]//*Proc of the International Conference on Computer Vision and Pattern Recognition*. San Francisco, CA, USA, 2010: 2046–2053.
- [12] Chapelle O, Haffner P, Vapnik V N. Support vector machines for histogram-based image classification [J]. *IEEE Transactions on Neural Networks*, 1999, **10**(5): 1055–1064.
- [13] Castrodad A, Sapiro G. Sparse modeling of human actions from motion imagery [J]. *International Journal of Computer Vision*, 2012, **100**(1): 1–15.
- [14] Michalis R, Iasonas K, Stefano S. Discovering discriminative action parts from mid-level video representations [C]//*Proc of the International Conference of Computer Vision and Pattern Recognition*. Rhode Island, USA, 2012: 1242–1249.
- [15] Sanin A, Sanderson C, Harandi M T, et al. Spatio-temporal covariance descriptors for action and gesture recognition [C]//*Proc of International conference on Application of Computer Vision Workshop*. Sydney, Australia, 2013: 103–110.

## 分层特征组的行为识别

周同驰 程 旭 李拟琚 徐勤军 周 琳 吴镇扬

(东南大学信息科学与工程学院, 南京 210096)

**摘要:**为提高视频人体行为识别的性能,提出了一种分层建模行为的方法.该分层模型根据人体运动的属性概述不同时空域的行为内容.首先,利用时间梯度并结合连贯的运动模式约束提取稳定、密集的运动特征作为点特征;然后,采用自适应尺度核的 mean-shift 聚类算法标定这些特征.具有同一标签的特征组通过最大池运算产生身体部分表示后,累积大尺度的视频体内视觉词响应作为视频对象的表示.在基准的 KTH 和 UCF-sports 行为数据库上,实验结果表明所提方法增强了行为特征的代表性和判别能力,同时提高了识别率.与其他相关文献相比,所提方法获得了优越的识别性能.

**关键词:**行为识别;连贯的运动模式;特征组;部位表示

**中图分类号:**TP391.4